


For Reference
Not to be taken
from this library

APR 0 4 2005

WITHDRAWN

The New Encyclopædia Britannica

Volume 20

MACROPÆDIA

Knowledge in Depth

WITHDRAWN

FOUNDED 1768
15TH EDITION

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

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

SAN BRUNO PUBLIC LIBRARY

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

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

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

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

© 2005
By Encyclopædia Britannica, Inc.

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

Copyright under International Copyright Union
All rights reserved.

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

Printed in U.S.A.

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

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

CONTENTS

1 Geomorphic Processes
27 German Literature
39 Germany
133 Globalization and Culture
138 Goethe
141 The Forms of Government: Their Historical Development
148 Government Finance
160 Graphic Design
169 Gravitation
178 Greece
205 Ancient Greek and Roman Civilizations
342 The Classical Greek Dramatists: Aeschylus, Sophocles, Euripides, and Aristophanes
352 Greek Literature
361 Biological Growth and Development
441 Guyana
447 Gymnosperms
466 The House of Habsburg
472 Hamburg
476 Harvey
479 Havana
483 Hebrew Literature
487 Hegel and Hegelianism
498 Heisenberg
500 Helmholtz
503 Heraldry
519 Hinduism
559 The Study of History
624 Hitler
629 The Holocaust
635 The Homeric Epics
639 Hong Kong
646 Horses and Horsemanship
656 Human Rights
665 Humanism
678 Hume
682 Humour and Wit
689 Hungarian Literature
693 Hungary
715 The Hydrosphere
732 Ice and Ice Formations
760 Iceland
768 Ideology
773 Immunity

CONTENTS

Geomorphic Processes

Geomorphic processes are all those physical and chemical processes that affect the surface features of the Earth. They include tectonic activity and surficial earth movements such as rockfalls and landslides. Geomorphic processes also involve weathering and the erosion and deposition of the resultant rock debris by streams, glaciers, and wind.

This article treats the mechanics and dynamics of these various natural geomorphic agents and processes. In addition, it discusses the ways in which humans promote their performance in altering landforms and near-surface features. The direct effects of human action on the physical environment are covered as well. For further information about the origins and evolution of landforms, see the articles CONTINENTAL LANDFORMS; PLATE TECTONICS; EARTHQUAKES; RIVERS; OCEANS; and ICE AND ICE FORMATIONS: *Glaciers*. (Ed.)

This article is divided into the following sections:

Physiographic effects of tectonism 1
  Recent tectonic movements 1
    Seismic movements
    Slow tectonic movements
  Effects on the Earth's surface 2
    Uplift and denudation
    Relief features
Weathering 3
  Processes involved in weathering 3
    Physical processes
    Chemical processes
  Controls and rates of weathering 5
Slope movements 5
  Factors producing slope movements 6
  Types of slope movements 6
    Rockfalls
    Creep and bulging
    Landslides and debris slides
    Slumps, earthflows, and debris avalanches
    Solifluction
  Characteristics of unstable slopes 9
Fluvial processes 9
  Entrainment and transport of sedimentary particles 9
  Materials transported by natural rivers 10
  Erosion and deposition in natural channels 10
    Sand and gravel bars
    Bed forms
    Scour and fill
  Fluvial processes in different environments 13
    Trout stream
    Arroyo
    Channels in humid-temperate regions
    Rivers in canyons
    Cold region rivers
Processes of glaciation 14
  Glacial erosion 14
  Glacial transport 15
  Glacial deposition 15
  Glacial loading and unloading 16
  Periglacial processes 17
Wind action 17
  Transportation of rock debris by wind 17
  Effects of wind transport 19
    Modification of particles
    Deflation of surfaces
    Abrasion by sandblasting
  Deposition by wind 20
    Sand dune formation and migration
    Role of vegetation
Physiographic effects of man 22
  Rise of human agency 22
  Effects of modern man 23
    Direct effects
    Indirect effects

Physiographic effects of tectonism

It has long been known that tectonism has profoundly influenced the physiography of the Earth. The Russian scientist M.V. Lomonosov stated in 1763 that the Earth's relief is formed by shaking of the Earth (earthquakes) or long-term sinking and raising of its surface. The German geographer Alexander von Humboldt suggested that mountain ridges were formed by upwelling of molten lava (volcanism) from the Earth's interior. This idea was derived from observation of volcanic activity in Central America. The French scientist Élie de Beaumont, however, thought that the Earth's relief features were principally attributable to gradual compression or shrinking of the Earth's crust because of slow cooling since its formation. In this view, tangential pressures in the crust caused folding and general tectonic movements.

The Austrian geologist Eduard Suess noted that rock deformation was particularly intensive in mountainous areas, and he came to the conclusion that a close link existed between mountain ranges and crustal uplift and between oceanic and other basins and tectonic sinking or depression. In general, scientists today discriminate between tectonic movements that are geographically restricted and involved in mountain building (orogenic movements) and those movements that are broader, less localized, and associated with the uplift or depression of plateaus and basins (epeirogenic movements). Both types of tectonism are responsible for the Earth's present relief features or physiography. Modern investigations have led to the theory that all tectonism is produced by the movement of great blocks of the Earth's crust relative to each other (see PLATE TECTONICS).

RECENT TECTONIC MOVEMENTS

Seismic movements. Direct observations and theoretical generalizations suggest a close connection between earthquakes, which periodically occur in different parts of the globe, and recent tectonic movements. Such movements are considered to be mechanical dislocations or faults within the crust. When faulting occurs, earthquakes are generated. The detection, recording, and analysis of the resulting seismic waves provides information on the intensity of an earthquake and its location—both where it occurred on the surface of the Earth and at what depth.

Tectonic movements of a seismic character are often called rapid because the physical consequences occur swiftly. The reasons for these tectonic movements cannot be directly observed because they occur at depths of hundreds of kilometres, but strong earthquakes generally manifest themselves in repeated disturbances of the Earth's crust through time and thus produce discernible physiographic effects at the surface.

Rapid tectonic movements frequently occur in the Soviet Union, where intensive earthquakes of destructive character take place several times every year; less intensive earthquakes are even more common. The earthquake epicentres are geographically localized and this provides the opportunity to carry out regional seismic investigations and to forecast the locations and the intensity of possible earthquakes. The greatest frequency of intensive, rapid tectonic movements has been established as characteristic of mountain regions—from North Africa, the Alps, Balkans, and Caucasus to the Tien Shan, Pamirs, and the Himalayas and around the margins of the Pacific Ocean (Alaska, the Aleutians, Coast Ranges, Andes, mountains of New Zealand and Indonesia, and the Japanese islands.

Mountain regions of intensive seismic activity

the Kurils, and Kamchatka). There are reasons to suppose that the main sources of rapid tectonic movements are the zones of deep faults that extend from the crust into the upper mantle (zone beneath the crust) of the Earth.

Slow tectonic movements. Slow tectonic movements of the Earth's surface were detected first in Scandinavia (the Baltic, or Fennoscandian, Shield, an area of ancient crystalline rocks) and subsequently in the Canadian Shield. In both regions dome-shaped, elevated, ancient shorelines and deposits of former lakes and seas were discovered. These phenomena have been explained in terms of a rather broad upward arching during the last 10,000 to 12,000 years because of constant melting of thick glacial ice cover and consequent rising caused by lightening of weight. The maximum uplift of Fennoscandia for this period was about 250–275 metres (820–900 feet) in the northern part of the Gulf of Bothnia, according to some authorities; uplift in the Hudson Bay region of Canada has attained nearly 300 metres (1,000 feet). Simple calculation shows that the average rate of such tectonic movements reaches two to three centimetres (0.8 to 1.2 inches) per year in the tops of arches, gradually decreasing toward their peripheral regions.

Slow tectonic uplift and subsidence with such characteristics are common almost everywhere. These slow tectonic movements were detected by precise geodetic measurements, and their average velocities ranged from tenths of millimetres to several millimetres per year for plains areas and 10 millimetres (0.4 inch) or more per year in mountain regions. Unfortunately, data are available for no more than 100 years. For this reason, it is not yet possible to define the general trends (or periodicity) of such movements, as is possible in regions of postglacial isostatic uplifts, where there are records or available data. Nevertheless, the main conclusion about recent tectonic mobility, supported by seismic phenomena, may at present be considered proved by numerous geodetic, oceanographic, and geomorphological investigations.

Recent horizontal movements of separate parts of the Earth's surface also have been detected with the help of precise geodetic measurements and geomorphic observations. The features of horizontal displacement along the Talass-Fergana Fault in Central Asia or more complex movements of separate points along the San Andreas Fault in California are examples. In both regions single horizontal displacements of several metres have been observed immediately after strong earthquakes.

Recent tectonic mobility must be considered proved, although not all aspects are yet understood. The evidence consists of orogenic zones that coincide with seismic phenomena, shield areas in which isostatic movements occur, and regional lines of horizontal dislocations. It is not yet possible, however, to make a precise evaluation of tectonic mobility through time and, in particular, to define the periodicity, if any, that is manifested. Further measurements and observations are needed for a full understanding of these phenomena.

EFFECTS ON THE EARTH'S SURFACE

Uplift and denudation. Modern concepts of the origins of the Earth's surface features are based on acceptance of the existence of a dynamic equilibrium between vertical tectonic movements (uplift and subsidence) and denudational processes, which encompass the erosion of rocks and accumulation of the resulting sedimentary debris.

Rates of continental denudation

The German geographer Albrecht Penck considered data on the volume of transported river sediments and concluded in 1894 that a layer 0.8 millimetre (0.03 inch) thick is annually transported to the sea by rivers. More recent calculations have shown that the solid sediment transported by rivers is alone sufficient to account for reduction of the Earth's surface by 0.09 millimetre (0.004 inch) per year. Other values, given by various scientists, range from 0.01 to 0.02 millimetre (0.0004 to 0.0008 inch) in lowlands to 0.6 to 0.8 millimetre (0.02 to 0.03 inch) per year in mountains. Values as great as 0.16 millimetre per year have been given for the United States.

It should be noted that the sediments transported by rivers make up only part of the general volume of all sediments that are subject to continental denudation. Some sediment is not carried to the sea at all but is retained within the area being eroded—on terraces, hillslopes, in local closed depressions, and elsewhere (see below). This material, which is not transported to the sea, is not included in the calculations on which the foregoing data are based. Hence, values of continental denudation are greater than stated.

The general correlation between recent tectonic movements and continental denudation may be considered an indication of dynamic stability of the Earth's surface. The stability is characterized by the following process: slow tectonic uplift of certain areas stimulates erosion and denudation, whereas other parts of the Earth's surface undergo tectonic subsidence, and the eroded material tends to accumulate in place, although part of it is transported to the sea in one form or another.

Data on the rate of development of a normal soil profile (sequence of vertical layers) under a natural vegetation cover are instructive in this regard. The use of radiometric measurements for defining the absolute age of humus in the recent soils has shown that a time period ranging from several hundred to 1,000 to 3,000 years (depending on the genetic type of a soil) is required for the formation of a normally developed soil. This means that the natural "growth" of soil thickness ranges from one to 10 millimetres per year. It shows that under the conditions of natural denudation, a well-regulated dynamic system exists. The three basic processes of the system—tectonic movements, soil formation, and continental denudation—tend essentially toward a state of balance. The middle component of the system, the soil-vegetation cover, serves as its main regulator.

Equilibrium is lost when the natural soil-vegetation cover is disturbed, whether by human activities or natural causes. Rather widespread phenomena of natural denudation are evident in mudflows, which form in mountains during rainstorms when there is an exceedingly high surface runoff. They are characteristic of many high mountain zones, where thick accumulations of loose sediments are present, and in arid regions, where the vegetation cover is sparse. Catastrophic displacements of ground masses on slopes (avalanches, landslides, and mudflows) are quite common in these areas. They are usually caused by earthquakes; and many interruptions of the soil-vegetation cover on mountain slopes are characteristic elements of the natural landscapes and must be considered a reflection of disturbed equilibrium.

Mudflows as an example of natural denudation

Relief features. The state of equilibrium between tectonic movements and the processes of denudation should provide for a stable, level character of the Earth's surface; it is clear, however, that such is not the case. Aside from extraordinary disturbance of the equilibrium by catastrophic movements of sediment, more gradual tectonic movements and denudation are not balanced in the strict sense and may be progressively altered and directed over time. Consideration of this balance—or lack of balance—has led to the distinction among three main categories of elements in the structure of the Earth's surface. The largest category consists of first-order relief features, which are elements of the so-called geotecture. They include crystalline massifs, continental platforms, oceanic hollows, and large mountain systems.

Positive and negative elements of lesser magnitude make up the second-order relief features that complicate the surface of continents and ocean floors; these elements are called the morphostructure. This group includes plateaus, uplands, lowlands, ridges, and similar features. Morphostructures may be described chiefly as large forms of relief that emerge as a result of the interaction of tectonic and denudational forces. Tectonic movements play the leading role in this interaction.

Small elements of relief of the third order include such features as river valleys, lake basins, and dune or karst (cavern) forms. They are sometimes called elements of morphosculpture. The origin of these morphosculptural elements is largely dependent upon denudational processes.

It has been suggested that all tectonic movements that are expressed in present-day relief should be called neo-

tectonics. In a paper delivered by a Soviet scientist in the 1940s this connotation emerged from considerations of the general amplitude and character of the most recent tectonic movements. A review map of these movements in what was then the Soviet Union served as a model for international compilations of this type. The main basis for the work was study of the thickness of Pliocene-Quaternary deposits (younger than 7,000,000 years) that accumulated in tectonic subsidence areas (depressions and basins) and of the amplitude of uplift of ancient surfaces in mountain ridges. In these instances, the obtained values of uplift and deposition amounted to hundreds of metres of relief over distances of several kilometres. (I.P.G.)

Weathering

In general terms, weathering may be defined as the disintegration or alteration in situ of rocks at and near the Earth's surface, within the range of temperatures that occur there. The distinction drawn between disintegration and alteration highlights the difference between physical and chemical processes. Disintegration involves a breakdown of the rock into its constituent minerals or particles, with no decay of any of the rock-forming minerals. Chemical weathering, on the other hand, implies the alteration of one or more of these minerals. The phrase in situ is not meant to suggest that there is no translocation of material within the rock being weathered. Redistribution and reorganization of various minerals are clearly at work in many vertical soil profiles, and these processes lead to the development of distinct horizons or zones and the lateral movement of minerals. The rock mass as a whole remains in place, however. Finally, the location of weathering processes at or near the ground surface, plus the range of temperatures indicated, differentiates these processes from metamorphism, which takes place either deep in the crust or at higher temperatures than are characteristic of weathering reactions.

Weathering and landforms

Weathering is an essential precursor to erosion and transportation. Weathering reduces rocks to particles and to a condition suitable for transportation. Hence, it is a key phase in the eventual production of new strata, as well as in the formation of the alluvial lowlands. Weathering also plays an important part in shaping scenery. It is not too much to say that weathering, by exploiting various weaknesses in the Earth's crust and thus preparing the way for erosional agencies such as rivers, glaciers, and wind-driven waves, plays a major role in determining the form of the land surface over much of the continental area of the globe. Upland and plain, ridge and valley, hill and lowland are in large part a reflection of structural contrasts brought about by weathering. In addition, pedogenic accumulations of minerals can reduce infiltration of water, cause heavy and rapid runoff, and thus induce floods. Iron pans in the soils of Exmoor, in southwestern England, were an important factor in the flooding that partially destroyed Lynmouth in 1952. On a large scale, silcrete and laterite, for instance, form the caprock in many plateaus and mesas in arid Australia. Even travertine deposited in riverbeds may protect the landscape to such an extent that the riverbed becomes more resistant than the surrounding areas and eventually comes to stand above the rest of the land surface. An example of such inversion of relief, with a winding, travertine-capped old river course now forming a distinct ridge, has been described from Arabia. Finally, numerous minor forms, including elegant flares, caverns, pits and etchings, are a result of weathering.

PROCESSES INVOLVED IN WEATHERING

Physical processes. Weathering processes are complex, and several processes generally act together to achieve rock disintegration and decay. The processes may for convenience be labeled physical (or mechanical), chemical, or biological, but such treatment tends to disguise the essential complexity and interconnection of weathering activities. Very few weathering processes can be observed in action; scientists must be content to see the results and try to infer from them what has taken place. Many variables affect weathering reactions, and laboratory experiments fail to reproduce either the complexity or the immense duration of geologic time, which together provide the only correct context in which to view these processes. As a result, accounts of weathering have been dominated by speculation and theorization, much of which, though it seems logical enough, fails to accord with the evidence of nature.

Thermal expansion and contraction

Thermal expansion and contraction, for example, have long been cited as a cause of rock disintegration, particularly in the tropical deserts where great extremes of temperature are experienced. Man-made fires were used as an aid in quarrying in ancient Egypt and in India; bush or forest fires certainly cause superficial rock flaking; and some stones subjected to the intense, though local, heat of campfires are rapidly split. Reports of loud cracking noises in tropical regions have been attributed to the expansion or contraction of rocks on heating or cooling. As a consequence, numerous types of weathering, including granular disintegration, spheroidal weathering, onion (or onionskin) weathering, and sheet structure, have been explained in terms of thermal expansion and cooling as described below.

Two related processes were involved. First, it was argued that, because many rocks consist of minerals that expand by differing amounts when subjected to the same temperature change, they would, on heating, expand at varied rates. As a result, daily heating and cooling would eventually loosen the cohesion between the rock-forming minerals and would thus cause the breakdown of the rock into small particles. The process described is termed granular disintegration. Second, because rocks are poor conductors of heat, those parts of a rock mass exposed at the surface would expand and contract as they are heated and cooled, while those parts below the surface should remain at a constant temperature and volume. It was supposed that stresses would be set up between the outer and inner zones, the former eventually becoming separated due to the development of fractures, and repetition of the process would cause the development of concentric layers or sheets of rock. This process resulted in spheroidal weathering, onion weathering, or sheet structure, depending on the thickness of the detached rock masses.

It now appears, however, that most field evidence points to contrary conclusions. Granular disintegration, spheroidal and onion weathering, and sheet structure have all been found deep beneath the Earth's surface, well beyond the range of the sun's heat; and even in the tropical deserts, which should be most suitable for solar heating effects, the evidence is contrary to that required by the insolation hypothesis. Observations in the Egyptian deserts indicate that rock surfaces known to have been exposed to the sun's rays for 42 centuries display no detectable sign of disintegration, whereas the same rock types buried nearby beneath the desert sand, where there is just a little moisture, show clear signs of decay. Furthermore, laboratory experiments suggest that heating and cooling alone either achieve little or work slowly, whereas heating and cooling in the presence of moisture produce almost immediate effects. Apart from the local effects of the ephemeral high temperatures achieved in bush fires, little weathering is today attributed to thermal expansion and contraction. The forms once related to this process now are generally attributed to contact with moisture.

Pressure-release hypothesis

Similar doubts relate to the interpretation of sheet structure, the massive (up to nine metres) slabs of rock that commonly are developed in crystalline rocks such as granite but that also occur in sedimentary strata. Though once considered an insolational (derived from the sun's heat) effect, they have for many years been accepted as a manifestation of offloading or pressure release. In fact, they often are called offloading joints. The pressure-release hypothesis is based on the following argument. Granites crystallize deep within the Earth's crust under conditions of high hydrostatic pressure, so that the very fact that granite is now exposed at the Earth's surface in itself proves that erosion of a considerable thickness of overlying material has occurred and that there has been a drop in vertical loading. During the erosional unloading, the granite tends to expand in response to the decreasing pressure. Such expansion is radial and upward, in the direction of least

stress; in detail, it is perpendicular to the land surface. The radial stress is relieved by the development of fractures that are tangential to the stress—that is, roughly parallel to the land surface: hence, the occurrence of arched or gently dipping joints that subdivide the granite into the massive slabs known as sheet structure. An essential feature of the hypothesis is that the form of the land surface predates and, in a broad sense, determines the geometry of the sheeting joints.

All joints are in a sense due to pressure release. All are near-surface features and presumably all disappear at depth. Moreover, many surfaces recently exposed from beneath ice masses display numerous joints that parallel the glacially eroded surfaces, and it is difficult to look beyond pressure release in explanation of these forms, which are, however, essentially superficial. But deep-seated fractures are another matter; sheet structure extends to at least 90 metres. Considerable evidence suggests that the offloading hypothesis is not everywhere applicable and that it should not be used unquestioningly. The very residuals, frequently inselbergs (island mountains), in which sheet structure is commonly displayed, are preserved because they are under compression, not under tension as demanded by the expansive pressure-release hypothesis. In some areas the dip of the sheet structures is opposed to the slope of the land, rather than parallel to it; sheet structure affects sedimentary and volcanic sequences that have never been deeply buried, and in some areas the ages of land surfaces and the sheet structure are the reverse of that implied by offloading. Faulting and lateral compression in the crust, residual from past or continuing earth movements, appear to offer an alternative explanation of sheet structure and to account more satisfactorily for the field evidence.

Thermal expansion and contraction, and the pressure-release hypothesis, have been considered at some length because they demonstrate how a too-ready acceptance of seemingly reasonable ideas can lead to incorrect or incomplete interpretations. Most accounts of weathering processes are similarly oversimplified, if not fallacious. Although some are more soundly based than others, none is completely understood, and it is better to work from evidence of what has taken place rather than from what presumably should happen in given circumstances.

Freeze–thaw cycles and effects

In many subarctic regions, outcrops of finely bedded rocks carry a veneer of platy debris. In these areas, where the vegetation blanket is thin and discontinuous, the common oscillations of temperature around the freezing point affect the superficial layers of rock. Water lodged in numerous cracks is frozen and expands, exerting a pressure sufficient to widen the fissures in which it rests. On thawing, the water rests lower in the now wider crack, until it again expands on further freezing. Such repeated expansion and contraction of contained water causes many rocks, especially those that are naturally well bedded, to break down into slabs and plates. Many close observations of strata, in regions where the temperature fluctuates about the freezing point, indicate the effectiveness of the freeze–thaw mechanism.

Crystal-growth pressures

Crystallization of such salts as sodium chloride and gypsum is also cited as a cause of rock disintegration, particularly in arid regions. There seems little doubt that the pressures exerted by crystal growth can rupture weakly cohesive rocks. Roofing tiles on the eastern (windward) side of Port Phillip Bay (Victoria, Australia) commonly are disintegrated by the crystallization of salts blown in by spray, and clays also can be broken up and disturbed. But whether salt crystallization can shatter such strongly cohesive rocks as fresh granite is problematical. Certainly salts are a common product of granite weathering and in arid climates efflorescences of sodium chloride can be seen on sheltered rock surfaces far distant from the coast. It seems likely that such expansion may at least contribute to the total weathering process. Some writers, however, considering the problem of weathered crystalline rocks in Antarctica, where, at present, there is never any moisture because of the consistent extreme cold, suggest that salt crystallization is responsible for the shattering of the rocks. It is, however, pertinent to ask what the source from which the salts crystallize is, if no moisture is available, and to point out that other common difficulties facing weathering studies seem to be involved: did the weathering observed occur in relation to the present climate or in a slightly different set of climatic circumstances? Is the weathering observed in Antarctica taking place now, or is it, as it were, inherited from a former, warmer and moister climate (in which case, freeze–thaw action could be invoked)?

The role of plant growth and organisms

There is no doubt that tree roots can force aside considerable blocks of rock and widen pre-existing joints during growth. Even the tiny roots, or hyphae, of lichens can penetrate along crystal boundaries and cleavages and can accomplish some physical disintegration. Burrowing animals such as rabbits and termites open up avenues for other agencies, particularly moisture. The burrows of earthworms long have been recognized as a significant element in the soils; worms penetrate about 120 centimetres (four feet) below the surface and they pass about four tons of soil per hectare, on average.

Chemical processes. The understanding of physical weathering processes is beset with difficulties, but the chemical reactions at work in the regolith are even more complex. Although it is possible to set down plausible reaction formulas, the fact is that most of these are so oversimplified as to be misleading. For purposes of exposition, however, it is convenient to isolate certain processes and their results.

Solution effects

No mineral is chemically inert, and many are to an appreciable extent soluble in water. Some, like rock salt, gypsum, and limestone, react strongly with water and either go into solution or form products that are soluble. Even quartz is to some extent soluble in water. Some minerals are more soluble in salt water than in fresh water: orthoclase, one of the feldspars and a common constituent of crystalline rocks, is 14 times more soluble. It is likely that solution is in many instances the first stage of chemical weathering. Solution also produces many distinct forms such as pits and fretting patterns, and its widespread significance is indicated by the vast quantities of material carried in solution by rivers.

Materials in solution are translocated vertically within the weathering profile (vertical section that exhibits weathering alterations) or may be carried great lateral distances and be precipitated far from their place of origin. Because of the translocation of minerals in solution (as well as of fine particles in solid state) within profiles, distinct horizons or pans rich in iron oxides, lime, silica, or gypsum may be developed. In reality, lateral accession is indistinguishable from vertical relocation, but whatever the precise source of material, salts are taken in solution from one part of the weathering profile and concentrated in another. Nodular or sporadic accumulations develop first, but these coalesce to form continuous, frequently massive, sheets. Because they tend to form tough, impermeable layers, they are called duricrusts. Bauxite is an example of such an accumulation; it essentially is an alumina-rich crust.

The cause of the localized precipitation of such minerals is not clear. In some instances, reaction with groundwater with contrasted chemical properties may be involved. In the case of calcrete, a lime-rich crust, it has been suggested that evaporation of groundwater may give rise to precipitation of lime at the upper fringe of the water table and near the land surface. On the other hand, plants markedly accumulate some minerals, and it has been suggested that silcrete, a silica-rich crust, for example, is fixed by plants and then added to the regolith when the plants die and decay. Whatever the mechanism involved in their formation, great sheets of laterite, calcrete, and silcrete have accumulated in various parts of the world.

Hydration and hydrolysis

Water and contained radicals and gases combine with various minerals to form new minerals, sometimes of volumes significantly different from those of the original. These processes are known as hydration (the addition of water) and hydrolysis (the addition of hydroxyl [OH ions]). Thus, iron readily combines with water and oxygen to form various hydrated iron oxides that are responsible for the yellow and red coloration of many weathered profiles. Orthoclase, a common constituent of acid crystalline rocks, as well as of some sediments, reacts with water and carbon dioxide to produce a clay, typically kaolin, and a

soluble salt and silica. All common rock-forming minerals except quartz are converted to clay minerals by chemical weathering, principally by hydration and hydrolysis. Thus mica is hydrated to hydrobiotite and vermiculite, and eventually to chlorite, but kaolinite and gibbsite are other common products of hydration. Hydration is important on its own account, but it also prepares mineral surfaces for oxidation and carbonation and generally enables ionic transfer to occur more readily.

Oxidation, or the formation of oxides, occurs in the aerated zone of soils, probably by interaction of the oxygen dissolved in water. Oxides are a common constituent of the regolith, and much takes place through the agency of bacteria that derive energy from the oxidation of iron and other elements. In waterlogged anaerobic sites bacteria can bring about chemical reduction (loss of oxygen). Thus, sulfates are reduced to sulfides, and organic material is reduced by fermenting bacteria.

Carbonation, or the reaction of carbonate or bicarbonate ions with minerals, is an important intermediate step in the weathering of such minerals as feldspars, and carbonic acid, though weak, is a potent solvent in nature.

Silicification and desilicification can cause the conversion of one type of clay to another. Thus, in tropical lands desilicification of micas gives rise to kaolin and iron oxide, or, if taken further, to bauxite (gibbsite) deposits.

As is the case with physical processes, considerable chemical weathering is attained with the aid of organic agencies. Humic acids, produced by the decay of organic matter, promote weathering generally. The hyphae of lichens can apparently extract certain radicals from minerals. Green algae in tidal pools raise the pH (acidity–alkalinity index) of the seawater during the day, and the emission of carbon dioxide (CO_2) by algae and invertebrates at night brings about the solution of carbonate. Humus in general helps conserve moisture in soils and hence aids weathering in a number of ways.

Although numerous individual processes have been proposed, the details of many are obscure; most are complex, and many open the way for the activities of others. Hence the complex of processes embraced by the term weathering is more effective than the total of its individual components. Water, acting both directly and indirectly as a solute and as a vehicle for various radicals, is undoubtedly the most important single factor in weathering, though biological agencies have increasingly been recognized as of great significance. But however achieved in detail, weathering produces a relatively thick mantle of debris.

CONTROLS AND RATES OF WEATHERING

Several factors control the type and rate of rock weathering.

Mineralogical composition. Of the common rock-forming minerals, the order of susceptibility to chemical attack is the same as the order of crystallization from magma (silicate melt that yields igneous rocks on cooling): olivine is most vulnerable, followed by plagioclase, biotite, potash feldspar, muscovite, and quartz. Thus a rock like quartzite, composed overwhelmingly of quartz particles and cement, will weather only slowly, while a basalt, on the other hand, rich in ferromagnesian minerals and plagioclase, will suffer rapid alteration.

Texture. In fine-grained rocks the total surface area of the constituent minerals is very great. These surfaces are quite prone to chemical attack, but such minerals tend to be closely interlocked and do not offer avenues for physical attack. Thus the coarse-grained rocks are most susceptible to such processes.

Fracture pattern. Fractures such as joints and faults are, if open, avenues of weathering readily exploitable by various environmental agencies, but particularly by water. Thus greatly shattered and fractured rock masses are much more rapidly weathered than are essentially monolithic masses.

Climate. From what has been said so far it is clear that many weathering processes achieve their optimal activity in specific climatic zones. Thus, freeze–thaw occurs and is effective in subarctic regions, but not in arctic areas where temperatures are too consistently low. Insolational heating and cooling effects are greatest in arid tropics. And chemical weathering is most important in the humid tropics, where moisture and humic acids are abundant and where temperatures are consistently high. Here the rate of chemical reactions is three times greater than in temperate regions, although a decrease in the viscosity of capillary water and its more rapid circulation in the regolith compensates for this to some extent.

Erosion and topography. Erosion may, by the removal of weathered debris, expose new rocks to weathering (renewal of weathering). On the other hand, stable conditions may encourage the development of deep weathering and thick regoliths. Steep slopes and high peaks are well drained, but valley floors, and particularly enclosed depressions, receive water and solubles. Hence, in poorly drained areas, waterlogging and reduction may be characteristic, or there may be marked precipitation of dissolved salts, depending on the prevailing climate.

Time. Prolonged weathering of rocks produces minerals different from those that result from brief reactions and removal of the debris. Reaction series resulting from continued hydration, silicification, or desilicification, for instance, are known and some of them have already been outlined.

Man. Quite apart from man-induced erosion, quarrying, and excavations of various types, man has stimulated weathering by his pollution of the air. The industrial release of sulfur has produced sulfuric acids in the air, which has affected not only man-made buildings but also natural rock exposures (see also below *Physiographic effects of man*).

Rates of weathering

Turning to rates of weathering, various isolated estimates have been made. Limestone tombstones in northwestern England are said to require 250 to 500 years to weather to a depth of 2.5 centimetres (one inch). After 45 years the ash falls associated with the 1883 eruption of Krakatoa displayed strong leaching of alkalis and some of silica. At Soufrière in the West Indies, soil development and reforestation on volcanic ash were "normal" after only 30 years. On the banks of the Murray River, Australia, weathering of a sandy limestone has varied between one and 30 centimetres (12 inches) per century.

Such estimates, though of interest, are all meaningless in a geological sense. So many factors are involved that great variations in the rate of weathering occur in the same area; aspect, rock type, exposure to erosion, and other factors have their effect. Even deep-weathering profiles, the initiation of which can be dated (as, for example, on lava flows of known age), are of little help. Is the weathering going on now, or did the whole profile develop in the past? Was the weathering achieved during a short time span when climatic conditions were especially suitable, or has it been going on steadily over a long period? It is not possible, at present, to be confident on these points, so general and imperfect is knowledge of Earth history and weathering processes.

(C.R.T./Ed.)

Slope movements

Very few landscapes are totally flat. Nearly all plains contain isolated hills rising from the main land surface, and most plateaus have valley-side slopes. The form of hillslopes, using the term for the sloping surface of both hills and valleys, consitutes a distinctive landscape feature. Of major importance in shaping hill-slopes are earth movements that occur under the influence of gravity when the stability of a slope is disturbed either by natural forces or by human interference. Earth movements of the most varied nature depend on the interaction of a number of factors, including the angle of slope, nature of materials, and time.

The great diversity of slope movements can be classified according to the mode and rate of movement, form of the surface of sliding, and the type of earth material moved. The most common downhill movement, almost imperceptible because of its small rate, is called creep, a category that comprises mass movements of a very wide scale, ranging from the creep of slope debris through outward bulging to long-term gravitational slides on mountainous slopes.

Another large group of slope movements includes landslides (landslips), which are rapid movements of earth materials separated from the underlying stationary part of the slope by a definite surface. When the movements occur along a predisposed surface where there is jointing or bedding, the slide is designated as a glide. If a free fall is involved in the movement of blocks of solid rock, the phenomenon is called a rockfall. Slides along newly formed curved surfaces are called slumps, and mass movements involving high-water content are called earthflows and mudflows. On steep mountainous slopes, torrential rains may produce debris avalanches. A special case of flowage and slipping is solifluction; *i.e.*, the movement of a thawed surface layer on the frozen substratum.

Significance to people

Slope movements, which may become a serious economic problem in their extent and recurrence because they often cause great damage to the property and life of people, can be an insurmountable hindrance to human activity. Many disastrous landslides and rockfalls are known to have destroyed whole towns and caused hundreds of deaths. Extensive depreciation of agricultural land, as well as of wooded areas, may also be caused by slope movements. Major slides result in the complete removal or extinction of forest growth; trees are uprooted or become dry. Highways and railways traversing areas susceptible to sliding are not infrequently interrupted by landslides, particularly when the stability of slopes is disturbed during construction. There are cases in which railway lines have had to be abandoned because of permanent danger of sliding movements and consequent high maintenance costs. Slope movements frequently produce serious difficulty during major engineering construction projects, such as tunnels and dams.

Adverse indirect effects of earth movements on slopes include clogging of valleys by landslides that impound temporary lakes that endanger the downstream reaches by flooding after the natural dam has collapsed. Sudden landslides and rockfalls along seashores also have very disastrous indirect effects. In the Norwegian fjords, for example, landslides often provoke high-water swells up to several tens of metres that are detrimental to the inhabited coast. (Q.Z./Ed.)

FACTORS PRODUCING SLOPE MOVEMENTS

The variety of landslide types reflects the diversity of factors that are responsible for their origin. Such factors include the character and structure of rocks, the angle of slope, the soil or debris cover, climatic and groundwater conditions, and time.

Debris cover on slopes is prone to downslope movements. As physical and chemical weathering disturb the cohesion of newly exposed rocks, further material for sliding is supplied. The stability of rocks is also impaired by chemical changes induced by percolating water.

Slopes composed of resistant permeable beds that are underlain by weak, incompetent, impermeable rocks, such as clays, are quite prone to sliding. The underlying clays become saturated with water and are squeezed out by the weight of the hard rocks above. In stratified rocks the contributing factor is the downslope dip of beds. If this slope is undercut by river erosion, then the stability of beds is disturbed, and slipping takes place.

The vegetation cover is also important. The roots of trees maintain stability by their mechanical effects and contribute to the drying of slopes by absorbing part of the groundwater. Deforestation of slopes impairs the water regime in the surface layers and facilitates erosion. The grass mat is gradually worn away so that weathering proceeds more intensively and produces free debris.

Earth movements on slopes are frequently induced by an increase of slope angle. This may be caused by natural or artificial interference; for example, by the undermining of the foot of a slope by stream erosion or by excavation. Exceptionally, the angle of a slope becomes steeper as a consequence of processes such as subsidence or uplift of the Earth's crust. The increase in slope gradient increases the shear stress within the rock mass, and this disturbs the equilibrium.

Tremors produced by earthquakes also affect the equilibrium of slopes by evoking a temporary change of stress. Some disastrous rockfalls in high-mountain areas (*e.g.*, Peru in 1970) are known to have been caused by earthquakes. In loess (fine-grained silts) and loose sands, shocks can disturb the intergranular bond and thus lower the shear strength. In water-saturated fine sands—those in which water occupies the pore space between all sand grains—and some kinds of clays (quick clays), shocks may result in a displacement or rotation of grains leading to a sudden liquefaction of soil.

Effects of earthquakes, groundwater, and climate

The conditions of slopes are greatly affected by groundwater. Flowing groundwater exerts pressure on soil particles that impairs the stability of slopes. In fine sand and silt, groundwater washes out fine particles, and the strength of the slope is weakened by the cavities that are formed. Moreover, soluble cement may be removed and, consequently, the cohesion of rock and the shear strength decreased. If the groundwater is under pressure it acts to uplift the overlying impermeable beds, thus decreasing the stability of slope.

The factors listed above often combine with the influence of climatic conditions, particularly the amount of precipitation and frost activity. Rain and ice meltwater, for example, penetrate into joints (rock fractures) producing hydrostatic pressure, and the increase in pore-water pressure in soils induces a decrease of shear strength.

It has been observed that, under certain climatic conditions, slope movements occur repeatedly in extremely humid years; measurements of rainfall amounts have confirmed that there is a direct relationship between precipitation and frequency of landslides. In what was once Czechoslovakia, the monthly rainfall for 70 years of record showed a striking coincidence with the recurrence of slope movements. Systematic examination of such records makes it possible to predict the renewal of slope movements in areas liable to recurrent sliding and to warn of imminent danger.

In clayey rocks, the deleterious effects of atmospheric water are heightened when the rainfall comes after a long dry period: clayey soils are desiccated and shrunken, and water easily penetrates deep into the fissures. The disturbance occurs on the lubricated layer in which the water accumulates.

Among the numerous factors inducing earth movements on slopes, that of time must not be omitted. As the agents change in the course of time, several phases of development occur. These range from the first signs of disturbance of the equilibrium to general loosening of the mass, which is then propelled into motion, travels downslope, and is gradually deposited.

TYPES OF SLOPE MOVEMENTS

Rockfalls. Rockfall refers to the abrupt movement of loosened blocks or complexes of solid rocks detached from rock walls. Rockfalls are distinguished by a very high velocity resulting from the free fall. Their size ranges from isolated stones to enormous masses of rock. The stones and blocks that fall build up talus, or debris fans, at the foot of mountain slopes that in some places coalesce into extensive aprons. The slopes of these fan-shaped deposits range from 25° to 40°, depending on shape and form of stone fragments. Floods often carry away the loose material of these fans and deposit it farther downslope.

Origin and occurrence of rockfalls

The origin of rockfalls depends on the morphology of the slopes and on the jointing and fracturing of rocks. Rockfalls are frequent in mountainous areas, particularly in valleys that have been overdeepened by glaciers. Factors contributing to the loosening of blocks are climatic conditions, chiefly weathering, wedging effects of freezing water in joints, hydrostatic pressure of water in open fissures, and pressure of growing roots. The movement can be triggered by the undercutting of steep slopes by erosion or excavation, by earthquakes, or, exceptionally, by thunderbolts.

Hundreds of rockfalls have been recorded from young mountain ranges, such as the Alps, Carpathians, Himalayas, Andes, and Rocky Mountains. One of the largest was the rockfall in the valley of Bartang River, in the Pamir Mountains in 1911. A rock mass of about 4,800,000,000

cubic metres (6,300,000,000 cubic yards) fell and dammed the valley, creating a lake 75 kilometres (47 miles) long and 262 metres (859 feet) deep. When large rock masses drop into a lake or fjord, dangerous huge waves flood the coast. In the Norwegian fjords numerous rockfalls have occurred, and catastrophic results were caused in part by the suddenness of the event.

If a rockfall involves an extremely large mass of rock, the freely falling body detached high from the mountain face may move downslope with a speed of up to 215 kilometres (135 miles) per hour; the blocks are shattered to rock fragments, and the mass movement takes on the character of a flow. Rock streams in the Alps and other high mountains are mostly explained in this way.

Rockfalls are also frequent on rocky shores of lakes and seas, as well as on steep concave banks of erosive river valleys. In these cases the rockfall is an important agency in modelling the cliffs and contributes to the recession of the coastline; the rocks are eroded by waves, the cliff becomes oversteepened, and the upper part of the wall collapses.

Rockfalls also can be responsible for the recession of waterfalls, particularly when hard rocks overlie less resistant beds. The existence of Niagara Falls, for instance, is threatened by large-scale rockfalls. In 1954 about 185,000 tons of rock collapsed because of undermining by water erosion.

Creep and bulging. The simplest form of creep is the slow, almost imperceptible downslope movement of soil particles and rock debris. Creeping of loose rock fragments is the result of a number of processes, particularly those related to climatic agencies. During the winter months movement is facilitated by the loosening of rock fragments and heaving by frost. Upon thawing in spring, the particles, which have been lifted at right angles to the slope, fall back vertically under the effect of gravity, so that they move a small distance downhill. The creep of slope debris can be compared to plastic deformation occurring in a snow bed on a mountain slope. The rate of movement is greatest near the surface and decreases downward. The expansion of stones by heat and shrinkage on cooling also contribute to the downslope movements, because they are not equal at the upslope and downslope sides of stones. Subordinate effects can be produced by plowing, cattle treading, burrowing by animals, and wedging by plant roots.

Clayey surface layers move slowly downhill by the action of plastic deformation. These movements do not usually develop a discrete slide surface but occur over a wider zone. They are limited to the surface layer, which does not surpass the depth of temperature and humidity effects. Though the deformations amount to only a few millimetres, during the long intervals embraced by geological time they appear as a steady creep of slope deposits.

Creep results in the bending of beds (Figure 1). Friction that is active between the creeping debris and the surface of the bedrock produces a gradual bending of the bed faces cropping out at the surface. The dragged-out and disrupted layers of the bedrock become part of slope deposits, thus increasing their thickness.

Deformation by bulging

In addition to surficial creep, a slow movement disturbing the equilibrium conditions of a slope takes place under suitable conditions at a greater depth below the surface. These earth movements are caused by the squeezing of soft rocks from beneath the more solid overlying rocks. In English literature this phenomenon is designated as bulging; in some countries the result is called a valley anticline. This process involves the plastic flowage of the underlying rocks along a system of surfaces of potential sliding. The instability of the slope is perceptible only during a longer time interval, when the minute deformations have reached measurable values. In the advanced stage these creep phenomena may grade into true landslides.

In some regions the squeezing out of soft rocks in the lower parts of the valley slopes is so widespread that it causes serious economic problems. Bulging was first described from an area of iron-ore mining near Northampton (central England), where valley sides are composed of solid Jurassic limestones and shales (144,000,000 to 208,000,000 years old) that overlie soft Lias clays. Although the beds are nearly horizontal, the near-surface beds of solid

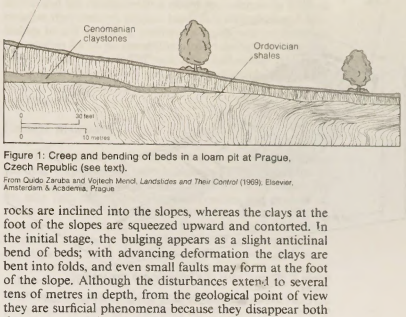


Figure 1: Creep and bending of beds in a loam pit at Prague, Czech Republic (see text).
From Quido Zaruba and Vojtech Mencl, *Landslides and Their Control* (1969), Elsevier, Amsterdam & Academia, Prague.

rocks are inclined into the slopes, whereas the clays at the foot of the slopes are squeezed upward and contorted. In the initial stage, the bulging appears as a slight anticlinal bend of beds; with advancing deformation the clays are bent into folds, and even small faults may form at the foot of the slope. Although the disturbances extend to several tens of metres in depth, from the geological point of view they are surficial phenomena because they disappear both downward and into the slope where the beds preserve a normal subhorizontal course. This deformation can be interpreted in terms of the upward squeezing of plastic substance from the loaded medium into the unloaded one. The stress that caused the heave of clays results from the difference in the loading of clays in the bottom of the valley and under its slopes.

Figure 2 shows an example of bulging in a river valley near Ostrava (Czech Republic). The valley occurs in marly shales of Cretaceous age (66,400,000 to 144,000,000 years old) that are pierced by sills of very firm volcanic rock called teschenite. The section shows that the teschenite body is broken into several blocks by a system of faults running roughly parallel to the valley. The soft shales that are squeezed out by the heavy rock blocks move toward the stream, which gradually carries them away. The main deformations probably occurred about 230,000 years ago, during the Pleistocene Epoch (10,000 to 2,500,000 years ago), because the steps between the blocks are filled with slope debris and loess loam.

The possible development of such deformations under present climatic conditions of the temperate zone has been the subject of much consideration. In general, squeezing of the substratum may occur whenever (1) the soft rocks within a limited area are released from the weight of the overlying rocks, and (2) this release gives rise to stresses, which, even if they do not surpass the shear resistance of plastic beds, may, in time, result in deformations.

Because the character of motion is essentially like that of creep, some long-term deformation of mountain slopes can be included in this group of earth movements. They consist of movement along planes of separation, such as planes of stratification, schistosity, or jointing.

Many analogous phenomena have been observed in young mountain ranges such as the Alps or Carpathians and occur most frequently on the slopes formed of phyllites, mica schist, and other metamorphic rocks that can be deformed by differential movements.

Movements related to creep can grade into sliding when a slide surface develops in the course of time and the movement is accelerated.

Landslides and debris slides. Landslides include a multiplicity of downslope movements, which, in contrast to

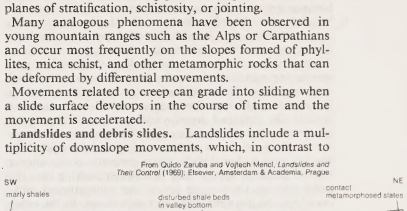


Figure 2: Squeezing-out of marly shales on the valley bottom of the Lucina River, Czech Republic.

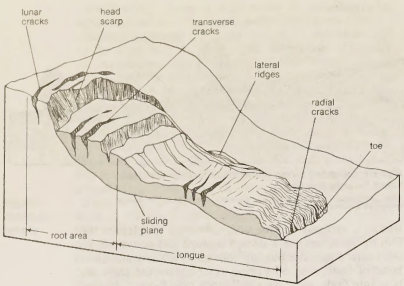


Figure 3: Principal parts of a landslide and characteristic cracks.
From Quido Zaruba and Vojtech Mencl, *Landslides and Their Control* (1969). Elsevier, Amsterdam & Academia, Prague

creep, occur along well-developed surfaces. A landslide occurs when the stability conditions of the slopes are disturbed either by the increase of shear stress imposed on the slope or by the decrease in shear strength of the rock building up the slope. The factors and processes that may provoke the change in the state of equilibrium have been mentioned previously.

Slide characteristics

The sliding movements involve bedrock or surficial deposits, but very often both bedrock and its cover are involved. The part of the slope that moves separates from the remaining mass along the plane of least resistance (Figure 3). This slip surface is commonly formed by a bedding, joint, or fault plane. The movements are generally rapid and originate when the planes of separation dip downslope and their continuity is disturbed at the foot of the slope. In stratified rocks with smooth, even bedding planes, the dip of beds is usually the maximum inclination at which the slope is permanently stable. If the beds are undercut by stream erosion, they maintain their position only by friction, which increases with the roughness and unevenness of the bedding planes. Friction can be reduced by climatic factors such as freezing and thawing of interstitial water or by hydrostatic pressure of water in the joints if the free outflow of water is obstructed. Failure can also be provoked by the increase of slope angle, as a result of uplift of the Earth's crust.

Rock slides on bedding planes or other surfaces of separation may be disastrous in mountain areas where, because of great height differences, the movement attains an acceleration nearly equal to that of a rockfall. In steep mountains the conditions are favourable for rock slides because streams with steep gradients cut readily into the bedrock, and the adjustment of the slopes cannot keep pace with erosion. The stability of slopes is particularly threatened when the dip is toward the valley.

Debris slides—movements of shallow slope debris and weathering materials above the bedrock—are categorized with landslides. Debris slides involve materials of only a few metres in thickness but may cover wide areas in some regions. On disturbed slopes, various stages of slipping are observable, from initial fissuring of the weathered layer up to advanced forms with several generations piled on top of one another. Debris slides generally occur after a heavy rainfall or spring thaw. Rainwater soaking into the soil lubricates the surface bed of the unweathered rock, on which the disrupted top layer slips down. In the other case, during the freezing of the ground the surface layers of debris are enriched by water rising by capillary action toward the surface from the lower unfrozen beds. During the spring thaw the water produces slaking of the surface layer and reduces the internal friction and the stability of slope.

Slumps, earthflows, and debris avalanches. The type of slope failure known as slump is common in homogeneous, poorly consolidated clayey rocks, such as clays, marls, claystones, and clayey shales. Slide-promoting forces are increased by the undermining of the slope by erosion or by excavation or by the overloading of its upper part.

Form and occurrence of slumps

Slumps have a characteristic form. The mass usually tears off along a concave head scarp and moves down a curved slip surface to accumulate at the foot of a slope, spreading laterally. Inside the sliding rock mass some minor scarps originate, so that it is broken into blocks that are tilted toward the slope. In the depressions of the hummocky surface, water accumulates into small lakelets, thus contributing to the instability of the slope. On either side, the mass is squeezed into longitudinal ridges that are often striated. The whole body is cut by cracks of different arrangement. Slumps may gradually increase by backward caving of the head scarp. The movement generally occurs along partial cylindrical surfaces, but the resulting slip surface is somewhat distorted. The depth and shape of a slump adapt to the geologic structure of the slope.

Slumps of large dimensions are frequent on river valley slopes and sea coasts. Along some river valleys deep slumps occur side by side, and the interlocking alcoves enlarge the slide area laterally. Slide material that weights the foot of the slope may help to restore the equilibrium of the slope, but tongues of the slumps are in most cases washed away during floods or by waves, and advancing erosion includes further movements. Typical examples of these slump movements are known from the coast of England near Folkestone (Figure 4) and from the valleys of various rivers (Volga, Moscow, and others) in Russia.

From Quido Zaruba and Bojtech Mencl, *Landslides and Their Control* (1969). Elsevier, Amsterdam & Academia, Prague

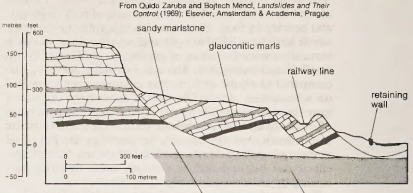


Figure 4: Landslide on the seashore near Folkestone, Kent.

On the lower parts of slopes, slumps frequently grade into flows, the shape of which is controlled by the topography. They move as sheets or streams that fill the erosion gullies or valleys. These movements—called earthflows, debris flows, or mudflows according to the material and consistency of the mass—represent a separate type of slope movement, one that differs from the slumps because of the higher water content that produces motion of flow type.

Role of water saturation

Earthflows generally originate in a large basin in the upper part of the slope, where debris and weathering material have accumulated. The movement is usually triggered by heavy rainfall. Interstitial water increases the weight of the debris cover and greatly reduces its shear strength. The loosened mass flows toward the foot of the slope, forming a loaf-shaped bulge. When it flows down a valley, it moves at a rate greater than is achieved over the slope surface, because water-saturated debris packed into a narrow tongue contacts the substratum over a smaller area; consequently, the friction is smaller. When all loose material has been emptied from the source area, the earthflow gradually becomes stabilized and overgrown with vegetation. If part of the debris remains there, heavy rainfall may provoke further movement, and the bulge at the toe will be overridden by younger material.

With increasing velocity earthflows grade into debris flows. Mountainous debris flows are sometimes called avalanches or *Muren*, a term currently used in the Alpine countries. As a rule, they originate above the timberline in gorges filled with rock detritus. During torrential rains, debris and larger stones travel at a tremendous speed down the stream channels. The material is unsorted, and

the ratio of solid particles to water is about 1:1. The rates of flow are so great that moving trains have been trapped and buried beneath debris avalanches. The Muren are very disastrous and often result from too great deforestation of mountain slopes. Scarcity of vegetation is also responsible for debris and mudflows in arid and semi-arid regions. This group of slope movements also includes volcanic mudflows. Volcanic explosions are usually accompanied by torrential rains that wash ash and ejected material downward as a mushy mass. Debris flows occasionally originate on glacier-covered dormant volcanoes. When volcanism recurs, the increase in heat before and during the eruption suddenly melts the glaciers, and the waters bring much debris downhill.

Solifluction. Solifluction is a combined flow and slip movement, which involves the surface layers on slopes in the subarctic region and, to a lesser extent, high mountain regions. The deeply frozen surface layers thaw to only a small depth during a short summer. Meltwater and precipitated water saturate the soil because they cannot percolate into the frozen, impermeable substratum, and the waterlogged bed flows downslope as a dense sludge. Solifluction may occur even on a rather moderate slope, because the soil moves readily on the frozen substratum.

Another somewhat special type of flow movement involves clay sediments of marine origin and is common in Scandinavia and Canada. These sensitive clays, called quick clays in Norway and Leda Clays in Canada, cover flattish areas situated about two hundred metres above sea level. The strength of these sediments progressively decreases because of the decrease of salts in the pore water. The ground and atmospheric water impoverish the salt content by osmotic processes. The decrease in salt concentration in pore water goes hand in hand with the decrease of bond between the clay particles and the bound water, and thus with the decrease in strength. Remarkably, the drop in strength is greatest toward the end of the process. The loss of strength results in a flow movement of unusually great rapidity. The failures of quick-clay slopes are treacherous because they may affect areas that are nearly flat.

Failure of quick clays

One of the largest failures of this type occurred near Verdalen, north of Trondheim, in Norway in 1893. A layer of sensitive clay was laid bare by stream erosion. The liquefied clay, with a volume of 55,000,000 cubic metres (72,000,000 cubic yards), flowed down in 30 minutes. The dense liquid covered an area of 8.5 square kilometres (3.3 square miles) and destroyed 22 farms. The flat terrace in the broad valley of the river Veddalselva is suggestive of absolute safety today, but the monument bearing the names of 111 people who were killed in 1893 testifies to the catastrophic possibilities.

CHARACTERISTICS OF UNSTABLE SLOPES

Slopes disturbed by sliding movements show a characteristic configuration: the head scarp is arcuate with a spoon-shaped depression, and the topography on the downslope side is hummocky and irregular. In active landslides the features are clear cut, whereas in the dormant, temporarily inactive landslides, the forms are effaced by rainwash and erosion or covered by vegetation. An important characteristic is the shape of the slope in cross section. Even a very ancient landslide is recognizable from its convex bulged toe made up of the accumulated slipped mass.

The growth of trees on unstable slopes reveals the presence and age of sliding movements. Trees, which on unstable ground become tilted downslope, tend to return to a vertical position during the period of rest so that the trunks are conspicuously bent. From the younger, vertically growing trunk segments, the date of the last earth movement can be inferred.

For distinguishing the slide areas, the presence of particular plants can be of help. Horsetail (especially *Equisetum maximum*) and coltsfoot (*Tussilago farfara*) are good indicators of slopes that are prone to sliding. Horsetails contain 50–60 percent silica and 19–30 percent potash in the ash. This high content suggests the presence of potassium and hydrated silicates in the soil and explains why horsetails thrive on sliding areas formed of potassium-rich (*e.g.*, glauconite-bearing) rocks.

The earth processes and factors that cause slope movements and the characteristics noted are intensively studied because they present serious hazards. Geologists and specialists in soil and rock mechanics endeavour to decipher the reasons for slope instability and to determine how fast and how far the loosened rock mass will move. Although some slope failures will remain unpredictable for a long time, at least those disasters caused by human interference can be avoided. (Q.Z.)

Fluvial processes

Over much of the world the reduction of mountains, the building of plains, and the sculpture of the landscape is brought about to a large degree by the flow of water. As the rain falls and collects in watercourses, the process of erosion not only degrades the land but the products of erosion become themselves the tools with which the rivers carve the valleys in which they flow. The process varies over time and from place to place. Materials eroded from one location are transported and deposited in another, only to be eroded and redeposited time and again before reaching the ocean. At successive locations the riverine plain and the river channel itself are products of the interaction of the mechanics of transport by the flow and the characteristics of the sediments brought down from the drainage basin above.

The fluid in a river is not pure water. Not always visible, the load of the river may be carried in solution, in suspension, or dragged along the bed. Solutes and particulate matter are both organic and inorganic. Neither the discharge of the water nor the related rates of erosion and deposition are constant in time or in space. Steep, narrow, rock-walled canyons may be excavated by corrosion of flowing water armed with abrasive particles aided by corrosion through chemical action. Elsewhere, sediments may be deposited to form broad alluvial fans, floodplains, or river deltas in lakes along the river course.

ENTRAINMENT AND TRANSPORT OF SEDIMENTARY PARTICLES

Erosion and transport of sedimentary particles is initiated when the drag, or shear stress, exerted on the boundaries of a natural channel by the flowing water is sufficient to detach a particle from the boundary. A particle of a given size and weight will begin to move when the shear stress exceeds a component fraction of the weight of the particle under water. In general, increasing flow accompanied by increasing velocity and shear stress results in progressive entrainment of particles from the bed and higher rates of transport.

There are essentially two distinct physical modes of transport of sediment. Bed load is that portion of the material in transport that is in continuous or partial contact with other particles on the bed; thus the weight of the moving particles is supported by contact with grains in the bed. In contrast, the suspended load is born up by the fluid eddies within the flow itself. Because the turbulent eddies both near the bed and in the fluid vary in intensity from moment to moment, the motion of the individual particles is highly erratic from place to place and moment to moment. In a statistical sense, however, at any distance above the bed, an equilibrium concentration is maintained with the number of particles settling through a given level balanced by an equal number of particles thrust upward by eddies. The concentration of suspended sediment at each level above the bed is a function of the settling velocity of the particles and the intensity of shear stress or turbulent exchange in the fluid at that level. Large particles, which settle most rapidly, are found near the bed, and progressively smaller particles are carried at greater distances above the bed. Similarly, where the particles are all of the same size, larger concentrations of suspended sediment will be found closer to the bed. If the particles are very fine, they may be nearly uniformly distributed with depth. The curves in Figure 5 show the way in which sediment concentration varies with depth for particles of different sizes or settling velocities at a constant condition of flow. Settling velocity is primarily a function of the size of the

Modes of sediment transport

Particle size and settling velocity

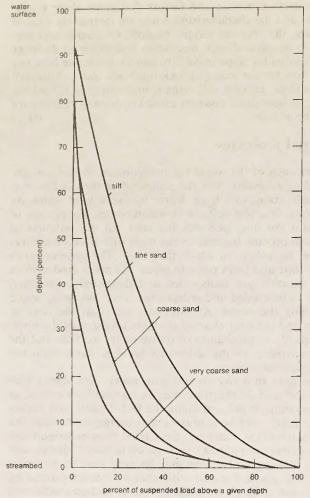


Figure 5: *Distribution of suspended load with depth for particles of different sizes.*
Coarse particles are concentrated closer to the bed, while finer particles, such as silt, are more uniformly distributed with depth.

particle. Assuming the availability of a range of particle sizes in a river, as the flow fluctuates and velocity or shear stress changes, particles transported near the bed at one time may be entrained and move in suspension at a higher flow. With diminution of the flow larger particles cannot be maintained in suspension and will settle to the bed.

Where the bed of a natural channel is comprised of noncohesive particles such as sand, as velocity and depth increase with increasing flow the particles in motion no longer behave solely as discrete individuals but instead the bed itself is deformed, resulting in the formation of sand ripples, dunes, and waves. As the rate of transport increases, these bed forms change progressively with changes in depth and velocity. The sand near the bed moves in a zone close to the dune or ripple surfaces. Usually, particles are carried up the backslope of the dune and deposited usually after sliding down on the steeply dipping face producing the downstream movement of the dune. Because deformation of the mobile boundary influences the flow distribution itself, the interaction between the flow, the boundary form, and sediment transport is complex. These interactions, and the progressive changes in bed forms observed in natural channels during passage of a flood, are discussed below, under *Erosion and deposition in natural channels.*

MATERIALS TRANSPORTED BY NATURAL RIVERS

The materials actually transported as a clastic or particulate load as well as those in solution in natural rivers are a reflection not only of physical laws governing the competence and capacity of the flow to transport material but also of the availability of materials themselves. For example, pure limestone terrain containing no insoluble materials would supply no sediment to streams draining the region. With the exception of occasional blocks broken off from limestone canyon walls, the rivers would carry only dissolved materials, and the stream beds would be devoid of sand, silt, and gravel bars. Such extremes, of course, are rare and, in general, rivers transport a mixture with coarser sediment in the bed load, finer particles in suspension, and ions in solution. Although these distinctions represent mechanisms of transport rather than sources of material, the sources are reflected in both the composition and relative abundance of each fraction.

The relative proportions of dissolved and particulate or clastic load as well as the distribution of the clastic fraction between bed load and suspended load are highly variable in nature (Table). In general, the percentage of clastic load increases as the climate becomes more arid. Less effective solution and weathering in arid regions apparently reduce the supply of dissolved solids to the stream channel system, whereas in more humid regions larger quantities of dissolved material are made available to streams. Topography, and particularly the composition of the bedrock, also can markedly influence the composition and quantity of the load. The Saline River in Kansas, for example, has a salinity of 1,000 parts per million by weight, a function of the saline deposits found in the watershed. Similarly, extreme concentrations of sediment have been measured in some rivers such as the Yellow River in China, where sediment comprised 40 percent of the total fluid, or tributaries of the Colorado River, in which 60 percent of the flow was sediment. In general, concentrations of suspended load in natural rivers vary from several hundred to 10,000 parts per million (1 percent) by weight. Some sand channels, such as the Loup River in western Nebraska, may transport 50 percent or more of the total load as bed load, but values of about 10 to 20 percent are probably more common.

EROSION AND DEPOSITION IN NATURAL CHANNELS

The supply and movement of sediment is intimately involved in the determination of the form and pattern of the river channel. A natural river flowing in sediments of its own making can maintain a stable configuration in two ways: first, by a balance of forces in which the drag of the fluid tending to erode the perimeter of the channel at any point is equalled or exceeded by the frictional or cohesive forces tending to resist the eroding force, and second, by maintenance of a rate of deposition equal to the rate of erosion. In contrast, a channel incised in rock is less free to adjust its form and pattern through deposition and erosion, but the bed of the channel will rise and fall with changes in the rate of transport of debris. The first case, in which a precise balance is maintained between the erosive force and the resistance of the boundary materials without any erosion, is relatively rare in nature. The banks of such a channel in noncohesive material are roughly parabolic and, if the flow is large, the channel cross section will consist of a wide central portion with a flat bottom and banks at each side roughly parabolic in form. Ideally, the maintenance of such an erosional equilibrium requires a

Equilibrium conditions and the migration of meanders

Dissolved and Suspended Load in Selected Rivers in Different Climatic Regions

river and location	drainage area (square miles)	average suspended load	average dissolved load	dissolved load as percent of total load
		(000,000 tons per year)		
Canadian River near Amarillo, Texas	29,700	6.41	0.12	1.8
Green River at Green River, Utah	40,500	19	2.5	12
Mississippi River, at the mouth	1,245,000	344	123	26
Delaware River, at Trenton, N.J.	12,300	1.0	0.83	45
Juniata River, near New Port, Pa.	3,354	0.32	0.57	64
Amazon River, at mouth, Brazil	2,722,000	499	242	33
Congo River, at mouth, Congo	1,425,000	31	98	76

delicate balance between the opposing forces, a relatively uniform material, and a constant flow, conditions reproducible in the laboratory but only approximated in nature.

A natural river channel in which the rate of erosion is balanced by the rate of deposition and the outflow of sediment to the reach equals the inflow can maintain a stable form while moving laterally across the alluvial plain. A meandering river is the most common illustration of this process. Erosion takes place on the outside of each bend near the point of maximum shear stress as a result of the curvature of the flow. Deposition on the opposite bank is associated with transverse flow near the bed and with slack water eddies adjacent to the thread of the current. Because the locus of erosion is downstream from the point of maximum curvature, progressive erosion in the downstream direction is associated with progressive deposition as the entire channel bend moves downstream.

Sand and gravel bars. Where the channel boundaries are straight, either because of the nice adjustment of discharge, gradient, and sediment, as is the case in some canals, or perhaps as a result of vegetation or channellization by man, sediment and sedimentary forms such as sand and gravel bars may move downstream in a progressive and orderly fashion. The movement of sediment and the configurations of the channel bed associated with such movement may be rather arbitrarily divided into three phases. First, channel bars may be deposited along the banks at sequential positions alternately on one side of the channel and the other. This configuration of alternating bars of gravel or sand in a straight channel is not unlike that which would be observed if the bends of a meander were "pulled out" to make a straight channel. Their spacing of roughly three to five channel widths appears to be related to the discharge and to the width of the channel. Second, sand and silt may move as dunes or ripples, a mode of transport determined by particle characteristics and the interaction of boundary form and the flow. A third mode of sedimentary deposit involves successive movement and deposition of discrete particles. Larger particles may move different distances depending upon their size, shape, and specific gravity. Boulders will move less frequently and, on the average, more slowly than smaller particles, producing a differential rate of downstream migration of particles of different sizes.

Accumulation and movement of gravel and sand in bars appear to resemble the movement of what is called a kinematic wave. Discrete particles do not move independently of one another but interact or interfere with each other in much the same way as automobiles on a highway. As a result of this interaction, the particles accumulate in groups or agglomerations that move downstream as "waves." The average downstream rate of movement is then represented not by the movement of the individual particles but rather by the average rate of movement of the group of particles constituting the wave. This wave phenomenon is similar to that observed on a highway crowded with automobiles, where, when the number of automobiles is low, the automobiles interact very little and each moves at its own rate. With an increase in the number of automobiles, however, interaction takes place and groups begin to form such that there are agglomerations and openings in successive positions along the highway. The celerity or rate of movement of these waves is determined then by the density of particles (or automobiles) in a given length of stream channel (or road), by the characteristics of the particles, and by the conditions of flow.

Flow and transport variation with time

Because the flow of water in natural channels is not constant, each mode of transport in a river also varies with time. The alternation of high and low flow in most of the rivers of the world not only influences the rate of transport but also the attendant forms of the channels themselves and the deposits associated with them. In the natural world the time scale of variations in flow may be matters of minutes, days, years, decades, or millennia. Peak flows from thunderstorm rainfall may occur in a matter of minutes after the start of heavy rain in streams or in urban rivers. Storms of longer duration may produce high water lasting for days or weeks. At another time scale are the seasonal or annual variations such as the cyclical rise and fall of the Nile each spring as a result of snow melt and rains in the headwaters, a variation common to many major river systems such as the Colorado, the Rio Grande, the Ganges, and the Yukon. Lastly, successions of dry years may be followed by wet ones. Such climatic variations are less periodic in occurrence and may encompass periods from decades to thousands of years in duration.

Variations in flow rate are associated with variations in transport. The dissolved load, responding to the characteristics of the source rocks as well as the flow, often decreases in concentration as a result of dilution of the salt concentration by the direct runoff from streams. In contrast, suspended load generally increases with increasing flow. Wash load (*i.e.*, materials such as clays and fine silts) may be readily removed from the watershed, for example, when spring rains follow the melting of snow and particles of soil are detached by cycles of freezing and thawing in early spring and late winter. The first spring rains readily remove the prepared materials to the streams. The transport of bed load also increases in response to increase in velocity and shear stress accompanying the passage of higher flows or floods. As in the theoretical or laboratory condition, the change in flow in the river channel is accompanied by a change not only in the concentrations of dissolved, suspended, and bed materials but by changes in the form of the bed itself as the flow increases in depth and velocity. Changes in the form of the bed associated with the passage of a high flow in a river are shown in Figure 6, whereas Figure 7 shows concurrent changes in flow, velocity, and depth.

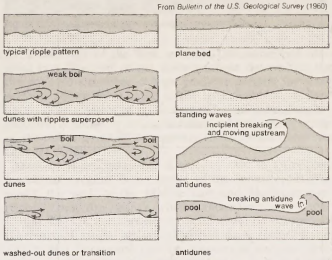


Figure 6: *Forms of bed roughness in alluvial channels.* This sequence of forms is related to the depth and velocity of flow and to the rate of sediment transport.

Bed forms. As noted earlier, after the initiation of movement, ripples and dunes develop on the bed. With increasing depth and velocity of flow in the channel, the dunes grow in size. Experimental and field observations indicate that the amplitude or height of the dune increases until the velocity of flow over the dune prevents further accumulation on the crest. At equilibrium, dunes may cover the bed in the same way that they do the windblown surface of a sandy desert. If velocity increases more rapidly than depth, a transition occurs, dunes begin to be erased, and all or part of the bed will be planar, with dunes covering the remainder. Beyond this transition, continuing increase in velocity leads to the formation of antidunes. Dune forms on the bed move upstream as a result of the displacement of material by scour at the upstream position while net transport continues in the downstream direction. The formation of antidunes is associated with the formation of standing waves on the water surface. Unlike dunes, the crests and troughs of antidunes are in phase with the crests and troughs of the surface waves.

Dunes

With declining flow in the channel and reduction in velocity and depth, the sequence is reversed. The bed is gradually transformed from the antidunes to the mixed and planar bed, to dunes, and thence to an irregular bed of sand if the flow declines to zero. A smooth bed is rarely if ever attained inasmuch as declining flow both erodes the dunes and deposits finer sediments on the channel bed.

Thus the forms remaining in a sand channel are usually dissected dunes and ripples.

Each of the successive configurations of the bed is associated with a particular set of flow conditions, and with changing discharge a complex interrelationship exists between the geometry of the boundary as defined by the configuration of the channel bed, the concentration of sediment, and the flow parameters such as velocity and depth. In some cases the same discharge on the rising stage of a flood may be associated with a high velocity and a relatively low depth, while on the declining stage the depth may be higher and the velocity lower (Figure 7). Resistance to flow is greater on the falling than on the rising stage. This difference is presumed to be due primarily to decline in flow so rapid that there is insufficient time for the transition of the bed forms. Dunes of larger amplitude associated with the higher flow remain at the lower stage and these larger forms provide a greater resistance.

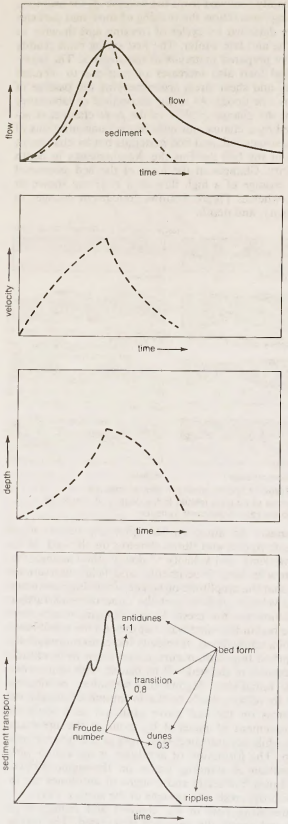


Figure 7. Sequential changes in flow (discharge), velocity, depth, and sediment transport associated with the passage of a flood in a natural channel with a sand bed. Changes in bed form and the approximate Froude number at which such changes might be observed are shown in association with changes in the rate of sediment transport.

The scale of dune forms may be exceedingly large. In a large river such as the Mississippi, at depths of flow from 21 to 27 metres (70 to 90 feet), dunes or dunelike forms of major proportions have been observed, with amplitudes from 5 to 12 metres and longitudinal spacing from crest to crest of 15 to 152 metres. Dunes can be seen in the beds of many rivers and creeks carrying sand, and dunes of cobbles and gravel occasionally have been observed after great floods.

Magnitude and frequency of flow events

Variations in flow produce variations not only in the rate of movement of particles of different sizes but also in the distances that these move in successive intervals of time. A large boulder in a small trout stream may be moved only by rare and very large floods. In contrast, material in solution is transported continuously as long as there is flow in the stream. Between these extremes lies a continuous range. Small particles are carried more often and for greater distances during each rise in the flow than are successively larger ones. Huge boulders may be moved a matter of centimetres or perhaps hundreds of metres by a flood that occurs on the average perhaps once in a hundred years. Sand may be transported continuously in a stream where the flow never declines below the point of incipient motion of the sand particles. Thus individual particles move in steps of varying lengths that depend upon the duration of the flow in which they are transported and their size, shape, and distribution.

Flood events of large magnitude but infrequent occurrence may move very large particles as well as large quantities of material but only for short periods of time. In contrast, more frequent events will move smaller sizes but over longer periods of time. In sand bed channels of ephemeral streams, such as those in the southwestern United States, most of the movement of both suspended and bed materials may occur during relatively infrequent events, which recur perhaps four or five times per year. These few floods not only mold the channel form, but they may transport 80 to 90 percent of the total annual load carried by the river. In contrast, where flow is perennial and of relatively large magnitude, more of the dissolved and suspended load may be transported by flows of modest magnitude throughout a large part of the year. Under such circumstances, the relative contribution of the large and infrequent flood event is thus of lesser significance. To the extent that generalization is possible, existing evidence suggests that the greater the variability of stream flow and the larger the particles to be carried, the more significant are the floods of large magnitude and infrequent occurrence. Similarly, where flow is less variable, dissolved load a significant proportion of the total load, and the suspended load composed of fine particles, more frequent events assume greater importance.

Scour and fill. In a given reach of channel, fluctuations in flow, accompanied by changes in the rate of transport of both suspended and bed material, produce a scouring or filling of the channel bed. In a straight channel this may amount to a few centimetres. In bends or narrow reaches, scour may be to depths of a metre or more and in rock-walled canyons as much as six to nine metres (20 to 30 feet). The cumulative effect in any given year may result in the temporary lowering of the stream bed. The bed may be lowered in some years and raised in the next or succeeding years. Over a period of time the average elevation will remain constant. A similar process of erosion and deposition characterizes the lateral or down-valley movement of the channel. In any one year erosion may exceed deposition, but over a period of years the pattern and form of the channel remain the same. Thus, the equilibrium referred to earlier constitutes an adjustment to the quantity of water and sediment delivered to the channel, an adjustment maintained within the natural variations in flow and sediment load normally experienced in a given climate or hydrologic environment. Equilibrium of the river channel then is associated with a constancy of climate viewed as an average over a period of years.

Changes in hydrologic conditions

A progressive increase in the amount of rainfall on the watershed, a change in the vegetative cover, or a combination of such factors, however, may produce a progressive change in the hydrologic conditions governing the river channel at any point. Such changes might be climatic, they might result from changes in land use or, for example, they might be brought about by construction of a dam and major reservoir. These long term or progressive changes are referred to as degradation or aggradation in contrast to the processes of scour and fill that encompass fluctuations of short duration. Changes associated with man-made works, in fact, provide a good illustration of the nature of the river response to changes in climate. A reservoir, by impounding flood flows for release during periods of low natural flows, alters the frequency distribution or pattern of river flows. In addition, the reservoir becomes a sediment trap, which reduces the quantity of sediment delivered to a channel below. A river previously in equilibrium with the natural flow and sediment supply is thus subjected to a new set of controls that, in turn, result in a succession of changes in the river channel itself. Thus, where scour took place in one year to be followed by deposition in succeeding years prior to construction of the dam, after construction the elimination of floods and sediment supply produces progressive lowering of the channel bed, or degradation, rather than an alternation of scour and fill (see also below *Physiographic effects of man*).

In contrast, progressive accumulation, or aggradation of material in channels, also results from changes in climate, from man-made works, or at the interface of land and water where sediments accumulate in deltas. In some locations, such as on the Yellow River, the river bed has risen to well above the surrounding countryside in the delta because successive floods deposited coarse sediment along the river banks and on the bed. Elsewhere, changes in climate have resulted in the delivery of large quantities of sediment to river channels, producing alluviation or filling for great distances over entire channel systems.

Rates of erosion and deposition

Rates of accumulation or degradation are highly variable. For over a thousand years the Nile has risen at an average rate of about one metre (more than three feet) per hundred years. In contrast, during the period of hydraulic gold mining in the Sierra Nevada the Sacramento River rose as much as three metres (10 feet) in 35 years. Valley fills or terraces along major rivers in the great plains of the western United States indicate that some rivers may have cut through the alluvial materials in their valleys at rates on the order of three metres or more in 100 years, although these estimates are crude at best. Below Hoover Dam on the Colorado River, the bed of the channel was lowered ten feet in five years after closure of the dam.

FLUVIAL PROCESSES IN DIFFERENT ENVIRONMENTS

The characteristics of a river and the processes associated with it are determined by a combination of the geology of a region and the climatic conditions responsible for providing the flow in the river. The term geology includes the underlying structure of the region, such as the presence or absence of mountains, as well as the composition and distribution of the bedrock. The latter, in turn, determine the quantity and often the size fractions of the materials delivered to the stream system. In contrast, although the climate may be influenced by the elevation of the region, major aspects of the atmospheric circulation will determine the amount of precipitation and seasonal variations in temperature. Indirectly these will affect the vegetation, characteristics of the soils, and the distribution of runoff or flow to the river system.

Clearly, innumerable combinations of geologic and climatic controls must exist in nature. Each segment or reach of a river will be dominated by a set of conditions of local origin as well as by a set determined from the larger area of the drainage basin upstream. Despite these obvious variations, however, some combinations of controls are sufficiently common to allow rather rough characterization of rivers in different regions. It must be emphasized, however, that the key to understanding why rivers look and behave as they do lies in a knowledge of the mutual interaction of the controlling factors such as discharge, sediment, and rock type. What might be called characteristic regional types simply represent particular combinations of these factors as well as the existence of a control in some regions or environment. Many complex factors are involved in developing a certain type of river or stream; each stream is classified according to the factors affecting it.

Trout stream. The so-called trout stream, or brook, for example, is indeed a recognizable type of river. It is usually characterized by a vigorous flow (of cool water), high gradient, and a bed of cobbles and boulders bordered by mossy banks and trees. Climate and elevation determine

Grant Heilman



Figure 8: A boulder-bed high mountain stream, Crandall Creek, Wyoming.

the availability of the flow, whereas the coarse bed material is a function of the steep gradient determined by the geologic structure. A high gradient permits the flow periodically to move some of the larger cobbles producing in time a series of pools and riffles. In some brooks, particularly those in coarse sand or cobbles, the spacing of the riffles or shallows appears to be proportional to channel width.

Arroyo. A second rather distinctive river in an opposing climatic regime is the arroyo of the Spanish-speaking world, the wadi of the Middle East and Mediterranean region. These dry washes, which experience periodic torrential flows, are otherwise dry sand channels comprised primarily of sands, perhaps dissected dune forms, some gravel, and occasional cobbles, depending on the local geology and topography. Because of the aridity and high temperature of the bed, vegetation is usually sparse. The combination of dry bed, absence of vegetation, and rapid changes in flow permit large variations in the configuration of the sand bed and in the amount of scour and fill. In addition, because such channels are often on relatively steep slopes and in noncohesive materials, they tend to be wide and shallow and subject to high velocities at shallow depths, a condition promoting the formation not only of dunes but of antidunes along with rapid rates of sand transport.

Dry washes

Channels in humid-temperate regions. In contrast to these rivers in semi-arid regions, the well-defined channels of the humid region are a response not only to the regularity of the flow but to the presence of fine sediments and vegetation that stabilize the river banks, emerging point bars, and other deposits laid down by the river. These channels, composed of silt banks stabilized with vegetation, are narrower and less mobile. Not infrequently floods that overtop the banks deposit on the adjacent plain varying amounts of silt, sand, and mud, thus creating a floodplain composed of channel and bar deposits overlain by finer "overbank" deposits. Where the channel is confined or maintains itself within a relatively narrow belt within which it meanders, the fine-grained deposits of floods may predominate in the floodplain.

The preceding descriptions suggest that in alluvial river channels flowing in sediments of their own making, there is a range of channel or river types, from the wandering channel that deposits bars over a vast shallow plain to the

well-fixed, stabilized, narrow channel bordered by vegetation. In the former the ability of the channel to alter its course at will by erosion and deposition, a condition common to large rivers flowing on an apron of debris deposited at the front of a glacier, creates a broad flat plain composed of sand and gravel bars deposited within the channel and in rapidly migrating bends, coupled with the overbank materials that are deposited from rapid rises in stage.

At the opposite extreme, the established channel in cohesive material bordered by vegetation will move laterally at a modest rate accompanied by deposition of sediments upon which vegetation will become established. With each flood above the normal bank height (a level attained on the order of once each year or two) sediments may be deposited over nearly the entire width of the valley. Where the river channel traverses the valley with rapidity, the bottomland is virtually all in channel deposits, while in the less mobile environment overbank or nonchannel deposits predominate.

Between the extremes, the wandering and stabilized channels, lies a spectrum of processes and resultant alluvial landforms, each dependent upon the relative rates of lateral migration of the river and of overbank deposition; that is, upon relative rates of erosion and sediment transport, and upon the frequency distribution of river flow. A given channel then may fall anywhere within this spectrum. The channel of the Missouri River along the boundary between northern Iowa and Nebraska, for example, contains many bars that may shift with each flood. At the same time, the river moves laterally in a broad meandering pattern experiencing progressive erosion as well as deposition over broad areas.

Rivers in canyons. Rivers in canyons such as the Colorado in the Grand Canyon, the Yangtze Gorge, or the limestone valleys on a smaller scale in Kentucky and Tennessee in the United States are confined by bedrock walls and valley floors. No longer free to alter either form or pattern save over long periods of time, the characteristics of the river are determined primarily by the rock itself and by slow erosional processes of abrasion and scour of rock material. Falls may occur where resistant beds are in contact with less resistant ones such as at Niagara, or long

Waterfalls

Josef Muench

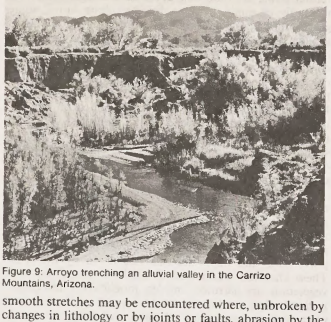


Figure 9: Arroyo trenching an alluvial valley in the Carrizo Mountains, Arizona.

smooth stretches may be encountered where, unbroken by changes in lithology or by joints or faults, abrasion by the river has produced smooth polished surfaces. Elsewhere, as in the bottom of the Colorado River, deep pools scoured by the flow may be separated by rocky rapids. In some places lateral abrasion excavates sheltered coves where rockwalls overhang the river itself. Rock canyon sections may occur in any climatic region where water has been or is currently available to scour the bedrock.

Cold region rivers. In extreme climates, such as in Arctic, Antarctic, or periglacial (cold, frozen ground) regions, the river may be frozen throughout much of the year. Here the river processes will be determined by the configuration of the bedrock geology and by the annual climatic regime, which determines both the way in which the river freezes with the onset of winter and the way in which the ice breaks up in the spring. Some rivers, such as the Yukon, break up throughout the length of the river in a relatively short period of time. The resultant breakup and the rising spring flood with which it is associated produce enormous forces that not only shatter the ice but cause it to accumulate in ice jams, to override the river banks, and to erode the channel banks as the high flow moves down the river.

This type of river presents quite a different pattern from that observed on other kinds of cold region rivers such as the Nelson River in Canada, which flows from Lake Winnipeg to Hudson Bay. The Nelson consists of a series of lakes and open water sections linked by bedrock falls or rapids. Freeze-up and breakup are discontinuous, beginning in the lakelike sections and controlled by the backing up of water above the controls at the rapids. The breakup of the ice is sequential and does not occur over the entire river, but instead the ice melts in the lakes and intervening backwater sections and, as the water rises, it begins to flow through and over the ice covering the falls. The flow scours areas adjacent to the falls themselves and gradually erodes through the ice capping the falls. The presence of ice and the movement of large ice blocks from the bedrock falls provide additional force to scour and to move boulders along the bed of the river. (M.G.W./Ed.)

Processes of glaciation

As a glacier moves over the land surface it modifies the terrain by removing material or by depositing debris carried within the ice mass or on the glacier surface. Each of these major glacial processes produces distinctive landforms that remain long after the glacier has disappeared (see CONTINENTAL LANDFORMS: *Glacial landforms*). In addition to direct glacial erosion and deposition, closely related geological events occur in nonglaciated regions as a direct consequence of the glacial ice, or the climatic conditions that caused the glaciers to grow and expand. These events include the depression of the Earth's crust resulting from the superimposed weight of glacier ice, the creation of lakes at the edge of the glacier during its retreat, the distribution of wind-transported sand dunes and silt blankets, emerged and submerged shorelines resulting from sea-level fluctuations in response to waxing and waning of the ice sheets, and the development of characteristic features produced by frost action in cold-dominated regions that were not covered by glacier ice.

GLACIAL EROSION

Glaciers are made up of interlocked grains of ice crystals. Laboratory experiments on single ice crystals and on samples of glacier ice show that ice deforms plastically when stress is applied. In nature the stress is gravity, the basic cause of ice movement. Two kinds of glacier movement have been identified: internal deformation of the ice by slippage or shear along certain planes in the individual ice crystals, and sliding of the base of the glacier on its bed. The velocity of the glacier thus consists of two components, an internal one resulting from deformation or creep of the glacier, and a basal one resulting from the sliding of the glacier on its bed.

The basal ice in many glaciers is at or near the pressure-melting point of ice, so that freezing and thawing occur at various times at the plane of contact between the glacier and the rock or soil beneath it. This permits loose debris such as silt, sand, cobbles, and larger rock fragments to be incorporated as part of the moving glacier mass. These basal, non-ice constituents of the glacier are abrasive tools that scratch, polish, or groove the rock or frozen ground over which the glacier moves.

The melting and refreezing of basal ice is a process that is capable of removing very large blocks of bedrock and incorporating them in the moving glacier mass. Many rocks of the Earth's crust characteristically contain intersecting fractures or cracks, called joints, which define incipient

Glacial quarrying

angular blocks that can be incorporated into the base of a glacier through the melting-freezing process. Large-scale excavation of many blocks (the individual dimensions of which may range from about a metre to more than three metres) is another form of glacial erosion known as glacial quarrying, or plucking.

The exact mechanism of glacial quarrying is not known because conditions at the base of a glacier cannot be directly observed. Tunnels, driven into glaciers along the ice-rock basal contact, have yielded very little information about quarrying. Generally, the excavation of rock by glacial quarrying seems to be restricted to rocks that are well-jointed. Massive, unjointed rocks, such as quartzite and others with few joints, are smoothed and polished by overriding glaciers. Glacially quarried surfaces, on the other hand, are rough and jagged because blocks were lifted out by the poorly understood quarrying process.

A good deal of controversy exists about the efficacy of glacial erosion. Some geologists admit that glaciers are capable of abrading rock surfaces but deny that glaciers produce any major landforms comparable to the large canyons cut by rivers. It is suggested instead that such features were produced by preglacial erosion and that glacial modification of them was insufficient to obliterate their preglacial forms.

The real question is not whether glaciers are capable of eroding the land surface but, rather, the magnitude of glacier erosion.

GLACIAL TRANSPORT

Material incorporated in a glacier by abrasion or plucking is transported by glacier flow until it is deposited directly by the ice, or until the glacier melts and leaves its load as a mantle over the landscape. Valley glaciers receive material from the valley walls confining them. Individual boulders and rock fragments, loosened by frost action from cliffs above the glacier surface, fall onto the glacier and are carried in conveyor-belt fashion as the ice stream flows down-valley. In some instances large landslides and debris flows may descend to the surface of a valley glacier, which then transports the rock and soil debris in a down-glacier direction.

Ablation and accumulation of glaciers

A glacier consists of two parts, the zone of accumulation and the zone of ablation. The former occurs in the upper reaches of the glacier where more snow falls each winter than is melted in the following summer. The latter is found at lower elevations on a glacier where, in addition to melting of the annual snowfall during the summer months, part of the glacier ice is destroyed by melting. If a glacier ends in a lake or the ocean, large masses of glacier ice become detached from the glacier terminus to become icebergs. Ablation thus is any process that removes ice from the glacier mass.

Rock debris that falls on a glacier in the zone of accumulation becomes buried by the snowfall of successive winters. A boulder that comes to rest on the glacier surface in the zone of accumulation has two components of movement as it is carried by the glacier. One component is down-valley, parallel to the general surface of the glacier. The other component is vertically downward as the boulder becomes covered by successive snow layers. Snow in the accumulation zone is transformed to glacier ice by a process of recrystallization as it is buried deeper and deeper each year. Debris falling on the glacier surface as airborne dust or boulders from a rockfall ultimately becomes imbedded in the glacier ice.

The flow lines of a glacier in the zone of accumulation are thus inclined downward toward the base of the glacier, whereas in the ablation zone, the flow lines have an upward component toward the glacier surface. Because of this pattern of flow, debris falling on the uppermost reaches of a glacier will be carried to the base of the glacier and then move upward until it reappears at the surface of the ice in the ablation zone.

The boundary between the ablation zone and the accumulation zone is called the snowline. It is best observed in late summer at the end of the ablation season. At that time the accumulation zone still has a residual layer of snow from the previous winter, whereas the ablation zone is strewn with debris released from the melting glacier ice.

If the amount of annual accumulation exceeds annual ablation over a period of years, then mass is added to the glacier. The glacier responds in two ways: it thickens, and it expands over a larger area. Valley glaciers expand in a down-valley direction and ice sheets expand by spreading outward in all directions. In both cases the glacier is said to be advancing.

If, on the other hand, ablation exceeds accumulation over a period of time, the glacier responds by thinning and by reducing its total area. The terminus of a valley glacier retreats in an up-valley direction while the edge of an ice sheet withdraws along the entire margin. During advance and retreat, the glacier continues to transport material in the direction of glacier flow.

From the foregoing it can be deduced that a vigorously advancing glacier may transport debris lying at its terminus by a "snowplow" action, thereby pushing material forward as the snout or margin advances. The snout is also the site of transportation of debris along thrust planes within the glacier itself. Thrust planes in some glaciers are exposed in crevasses near the glacier terminus or on the ablation surface, where the edges of the thrust planes occur as a series of parallel or subparallel bands emphasized by a concentration of dirt and stones that have been carried to the glacier surface by thrusting.

GLACIAL DEPOSITION

Material transported by glacier ice eventually comes to rest on the land surface. Glaciers that terminate in the ocean (tidewater glaciers), such as those discharging from Greenland, Alaska, or Antarctica, deposit material on the sea floor. If the terminus of a tidewater glacier is grounded on the sea floor, deposition is concentrated around the edges of the grounded portion as debris frozen in the glacier is released by melting. If the terminus is afloat, sediment released from the underside of the glacier will fall to the sea floor as melting progresses. These glacio-marine sediments are distinguished from other marine deposits by their heterogeneous detrital content.

Fluvioglacial processes at the ice margin

During the height of the ablation season, the terminus of a glacier on land is a chaotic mixture of melting ice, running water, and mounds of rock and soil debris recently released from the ice by ablation. These conditions change hourly; in late afternoon the discharge of glacial meltwater reaches a maximum, and large volumes of water can be seen flowing on the glacier surface in ice channels or debouching from tunnels in the ice. The flow of meltwater slackens after sunset, and by the early morning hours of the following day, the volume of meltwater being discharged from the glacier may be only a small fraction of the maximum flow of the preceding day.

These conditions give rise to a great variety of glacially

By courtesy of James Zumberge, University of Arizona, Tucson

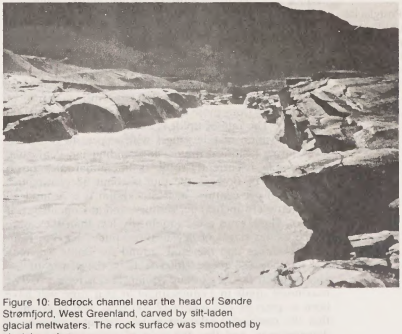


Figure 10: Bedrock channel near the head of Søndre Strømfjord, West Greenland, carved by silt-laden glacial meltwaters. The rock surface was smoothed by glacial erosion.

derived sediments ranging from ablation debris blanketed over stagnant ice masses to sands and gravels deposited by glacial meltwaters a few miles down-valley from the glacier terminus. The extremely variable conditions of transportation in space and time give rise to extremely varied deposition of sediments. Ablation boulders and unsorted rock rubble may slump onto a deposit of well-sorted gravels, or boulders may fall from the ice front into a quiet meltwater pool where fine sand and silt are accumulating on the bottom.

If a glacier margin retreats year after year, these chaotic ice-marginal deposits will persist over a wide expanse of recently deglaciated terrain. Pauses in the retreat may result in a concentration of the ice-marginal deposits in the terminal zone. If the ice front advances, it may override the older debris-covered areas or push some of it into ridges and mounds that remain for centuries after the glacier has retreated or disappeared entirely.

Meltwater issuing from the ice in rivulets on the ice, or torrential discharges from ice tunnels, generally converges to form a single channel or series of channels that divide and recombine in an anastomosing (braided) pattern. These braided rivers are choked with glacially derived sediment called outwash or glaciofluvial sediments.

If a glaciofluvial river empties into a lake, the coarser fraction of the suspended load is deposited at the river mouth while the silt and clay, called rock flour, are deposited on the lake floor. Glacial meltwaters heavily charged with rock flour are distinctively grayish-white in colour and are referred to as *Gletschermilch* in the Alpine regions of Europe.

GLACIAL LOADING AND UNLOADING

The Earth's crust tends toward a condition of balance or isostatic equilibrium. If the crust is considered to be a series of large blocks extending to a constant depth, approximately 20 to 50 kilometres (12 to 30 miles), each of the blocks will stand at a different elevation above sea level because of its difference in density. This condition of crustal balance is constantly changing because material from the higher-standing blocks (the continents) is being eroded and transported to the lower blocks (the ocean basins).

If glacier ice accumulates to some appreciable thickness over a large part of the Earth, the crust will be depressed about 300 metres (1,000 feet) for every 900 metres (3,000 feet) of ice. Crustal downwarping increases in magnitude from the marginal areas of a large ice sheet toward the zone of greatest thickness. Where the ice is thin, the crustal downwarping is zero, and where the ice is thickest, the depression of the crust is greatest.

Deformation of the crust because of glacial loading is not a permanent condition. After an ice sheet has melted away, the crust restores itself to a new condition of isostatic balance by rising in response to the glacial unloading.

Postglacial uplift in progress today

Crustal downwarping from glacial loading, and crustal uplift from glacial unloading are operative on the Earth today. The crust beneath Greenland and the Antarctic ice sheets is downwarped several thousand feet because of the load of glacier ice on each. In contrast, crustal upwarping is going on in Scandinavia, North America, and the British Isles as a result of the disappearance of the ice sheets from those areas somewhat less than 10,000 years ago. The rate of uplift is determined by the change in the elevation of tide gauges with respect to mean sea level during historical time. The Scandinavian Peninsula, for example, is currently rising at differential rates: north of the Gulf of Bothnia the rate is about 88 centimetres (35 inches) per century, near Stockholm it is around 40 centimetres (16 inches) per century, and at Copenhagen it is zero. These rates are undoubtedly less than they were in the earlier stages of deglaciation, but it appears from other evidence that the land around the Baltic Sea may rise another 180 metres (600 feet) in the centre of the area of uplift before isostatic equilibrium is regained. The maximum uplift in Fennoscandia at that time will have been as great as 760 metres (2,500 feet). If it is assumed that the postglacial rebound equals the total amount of downwarping, and that the ratio of downwarping to ice thickness is about one to three, then the maximum ice thickness for the Fennoscandian region was about 2,300 metres (7,500 feet). This is comparable to the measured ice thickness of the Greenland Ice Sheet, which covers about the same area today as did the Scandinavian Ice Sheet during the Pleistocene.

The depression of the Earth's crust by glacier ice produces some side effects in the nearby nonglaciated areas. Coastal regions are submerged, causing invasion of the land by the sea. River mouths are inundated by seawater and river gradients are reduced, giving rise to deposition of river sediments.

The postglacial rise of a previously glaciated land area is recognized in coastal regions such as Scandinavia, the eastern United States, and the Great Lakes region of southern Canada (Figure 11). The evidence comes from differen-

Adapted from Wayne and Zumberge, "Pleistocene Geology of Indiana and Michigan," in *The Quaternary of the United States*, by H.E. Wright, Jr., and David G. Frey (copyright © 1965 by Princeton University Press); fig. 4, p. 70; reprinted by permission

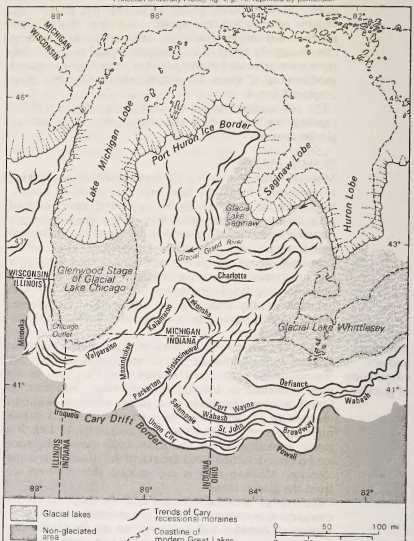


Figure 11: Upper Great Lakes region during Port Huron maximum, about 13,000 years ago.

tially uplifted or warped strandlines, which marked the shorelines of water bodies marginal to the retreating Pleistocene ice sheets. These strandlines, represented by beach ridge deposits, were horizontal when formed, but because of glacial rebound since deglaciation these linear features now rise toward the direction of greatest ice thickness. Accurate dating of successively younger strandlines provides the basis for determining postglacial rates of uplift. Crustal recovery patterns are depicted in maps showing isobases, lines connecting points of equal uplift in a given region, as around the Great Lakes or the Gulf of Bothnia and the Baltic Sea. The limit of uplift is defined by the zero isobase (no uplift) or "hinge line." In the Great Lakes region the hinge line has an east-west trend roughly from Milwaukee through Cleveland to New York City. South of that line, all strandlines of the ancestral Great Lakes are horizontal, a fact that implies no uplift. Northward from the zero isobase of the Great Lakes region, uplift is still in progress, but the rate appears to be declining. It has

been shown that the maximum uplift north of the Great Lakes occurred at Golfe de Richmond at about latitude 56° N on the east coast of Hudson Bay. There, at least 300 metres of uplift has occurred since the load of the ice sheet was reduced by thinning at the end of Pleistocene time. Assuming that complete uplift represents one-third of the ice thickness, the uplift at Golfe de Richmond reflects only about 900 metres of ice. This value is too small in relation to maximum Pleistocene ice thickness for the Greenland and Antarctic ice sheets, at comparable distances from their margins. It must therefore be assumed that uplift in the Hudson Bay region is still in progress, and that as much as 600 metres of additional uplift can be expected to occur.

PERIGLACIAL PROCESSES

As originally used, the term periglacial was an adjective used to describe the climatic conditions prevailing in a nonglacial zone marginal to a large ice sheet. As used today, the term is synonymous with "cold-dominated" and is applicable to a rigorous climate in which the freeze–thaw process predominates. Generally speaking, the periglacial environment is characterized by perennially frozen ground (permafrost), low precipitation, strong winds, and vegetation ranging from the boreal forests (central Alaska), through taiga (northern Canada) and tundra (Arctic Alaska and Siberia), to vegetation-free areas (northern coast of Greenland and glacier-free areas of Antarctica).

Periglacial environment

The periglacial environment was undoubtedly more widespread during the Pleistocene than now, and many effects of this environment are preserved in various landforms that resulted from periglacial processes that are no longer active. The landforms resulting from these processes are discussed in CONTINENTAL LANDFORMS. Whereas such features are not, in the strictest sense, landforms produced by glaciation, they are closely related to glacial climates and are commonly superimposed on terrains that originated through glaciation. Moreover, the presence of underground ice, and its thawing and refreezing are basic requirements for many of the landforms associated with the periglacial environment. (J.H.Z.)

Wind action

The term wind action embraces all aspects of the interaction between wind and rock and mineral materials on the Earth's surface. Wind is geologically most effective in areas where it acts with the greatest force for the longest time upon materials of greatest susceptibility. These conditions are attained more fully in dry than in humid regions. Consequently, the magnitude, abundance, and diversity of geological products of wind action are greatest in the principal desert areas of the world: the Arabian Peninsula, South West Africa, the Sahara, northwest China, Australia, the west coast of South America, and southwestern North America. Many of the large-scale patterns in arid regions visible from the air are the product of geological work of the wind. Local enclaves also exist within humid regions, where wind action can be a significant or even dominant geological process. Examples are sea- and lakeshores and the alluvial plains of rivers with variable flow. Many areas peripheral to glaciers are not arid in the usual sense, yet some of the most intense and effective wind action on geological record has occurred in such settings.

Wind displays unusual diversity as a transporting agent. It is one of the few agents that carries material uphill, and sometimes it works in opposition to other transporting processes. Streams of water, for example, carry sand from the land to the sea, where it is redistributed along shore in the form of beaches; in places wind picks up this beach sand and carries it back inland toward the source from which it came. In total weight of material transported, wind cannot match running water, but in terms of distance of transport it surpasses all other agents for materials as fine as dust. No part of the planetary surface escapes the influence of wind action; even deep-sea oozes contain wind-borne particles, and bits of snow are widely drifted by wind across the heart of the Antarctic Ice Sheet, where liquid water is unknown.

Geological activities of wind impinge upon the affairs and concerns of man. Deposits of windblown sand encroach upon railroads, highways, canals, airports, settlements, and other man-made works to the extent that sand control becomes necessary. Dust- and sandstorms temporarily close down major arteries of travel. Drought and man's disturbance of the ground permit wind to render areas of settlement and cultivation untenable. On the other hand, some of the richest croplands of the world occur on soils derived from deposits of windblown dust (loess), and wind has created aesthetically pleasing features such as the graceful form of lines and features of sand dunes.

TRANSPORTATION OF ROCK DEBRIS BY WIND

The flow of air. Because air has a very low viscosity, about one-sixtieth that of water, any movement or airflow at velocities greater than three or four kilometres (two to three miles) per hour is predominantly turbulent. A breeze of that velocity is barely perceptible to a standing observer. In turbulent flow, individual parcels of air within the flowing mass move in all directions, up, down, and sidewise, as the air moves forward. The other mode of movement is laminar flow, in which adjacent sheets of air slip past each other like playing cards in a deck, in essentially parallel paths without mixing. Because air is a perfectly viscous fluid, a thin zone of laminar flow always exists adjacent to any fixed surface, and an additional very thin film of dead air (air with no movement at all) lies in immediate contact with the surface.

Turbulent flow and the boundary layer

Within air moving across a surface there is a contact zone in which the velocity of translation changes from zero at the surface to the prevailing value at some distance therefrom. This is called the boundary layer, and above a natural ground surface its thickness is usually between one and 10 centimetres (0.4 and four inches). The boundary layer is a zone of great velocity change (gradient), strong shearing action owing to large differences in velocity at different levels, and unusually high turbulence. Its dimension and character are strongly influenced by the roughness of the ground. Wind velocity refers to the rate of movement parallel to the ground, and the prevailing velocity is that value measured at one to several metres above the surface.

Significant geological work (that is, wind action on rock materials) does not occur at prevailing velocities below about 16 kilometres per hour (10 miles per hour). Winds above this velocity are termed effective, and their capacity for doing work varies roughly as the cube of the velocity. If the velocity doubles, the capacity to do work increases eightfold. Strong storm winds, even though short-lived, therefore, can produce major effects, and not infrequently they cancel out the work of transportation accomplished by much more frequent but gentler winds.

Particle size and wind velocity. Wind-tunnel experiments show that different wind velocities are required to initiate movement of particles of different size, but of similar shape and density. The relationship is not linear, however. The highest velocities are required for the largest and for the smallest particles. A particle of some

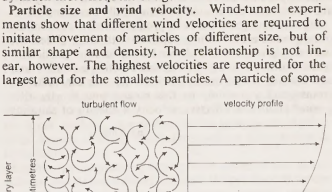


Figure 12: Diagrammatic view of the laminar and turbulent flow fields and the velocity distribution of the wind above the ground surface. The smallest particles are not lifted because they do not project above the laminar flow zone. The largest particles are subject to turbulent lifting forces but are too heavy to be moved. The "ideal particle," as shown, is that which will be transported by wind action.

intermediate size is the one that can be moved with the lowest wind velocity. This is the so-called ideal particle. On Earth it has a diameter of 0.08–0.1 millimetre (0.004 inch), and it starts to move at a wind velocity of about 16 kilometres per hour. On a planet such as Mars, however, where atmospheric density is less than 0.01 and gravity is about 0.4 that of Earth, the ideal particle size is about 0.7 millimetre (0.03 inch), and the required wind velocity is in the neighbourhood of 320 kilometres per hour (200 miles per hour). The reason small particles are hard to move is illustrated by Figure 12; they simply do not project high enough into the boundary layer for the wind to get hold of them.

Saltation of grains

Mechanics of movement. Wind moves particles by rolling or sliding (traction), by hopping (saltation), by suspension, and by a process of impact creep wherein the impacted grains slide or roll across the ground. Unless disturbed by other means, the initial movement as the wind attains an effective velocity is by rolling. As wind velocity and particle movement increase, the rolling grain may hit an irregularity and make a little hop. This gives the wind a better grip, and it accelerates the grain's movement. If the grain hits the surface in a favourable spot and fashion at the end of its leap, it rebounds to greater heights, the wind accelerates it even more, and the hopping process is magnified. This is saltation (Figure 13). Creeping grains travel at a rate of several millimetres per second but saltating grains cover many metres per second; the effects in terms of transport and of wear and tear upon themselves and other objects are multifold.

The height attained by a saltating grain depends upon its velocity, the rate and direction of its spin, its size, and the orientation and physical properties of the surface upon which it impacts. At high wind velocities (50 kilometres, or 30 miles, per hour), over resilient surfaces, individual saltating grains rebound to heights as great as three to six metres (10 to 20 feet). Most grains travel at much lower heights, about seven to 50 centimetres (three to 20 inches), however.

The distance covered in a single hop is governed by the above considerations controlling the height of rebound, and it ranges from a few centimetres to 10 metres or more. The mean distance over sand-mantled surfaces is on the order of 10 to 15 centimetres (four to six inches). Slow-motion photography and laboratory experiments show that most saltating grains moving under winds of 30–50 kilometres (20–30 miles) per hour approach the ground at angles between 10 and 16 degrees. For a grain rebounding to a height of one metre this means a hopping distance of three to five metres. Slow-motion photographs also record many midair collisions among grains. Such impacts produce abrupt changes of path, and some grains are returned to the ground at abnormally steep angles.

One saltating grain tends to produce others, particularly if the ground is abundantly supplied with sand. Unless the incoming grain scores a direct hit on another particle, it splashes down among other grains creating a small crater and blasting a half dozen particles into the air. Saltating grains quickly multiply by this means, and shortly after a wind attains an effective velocity a curtain of saltating grains becomes visible over surfaces supplied with loose sand. Over nonresilient surfaces, such as dunes, this curtain attains a height of only 30 to 45 centimetres (12 to 18 inches). Over resilient surfaces, such as stony plains or concrete highways, it easily attains heights of two metres (six feet) and more (Figure 13). Saltating sand grains seldom rise above knee height on the windward side of a sand dune, even in strong winds of 50–80 kilometres (30–50 miles) per hour. Under corresponding conditions over a paved highway, humans and vehicles both suffer.

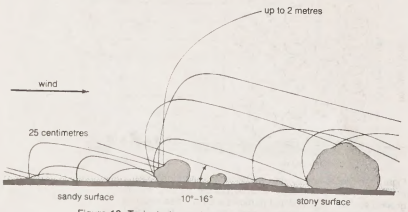


Figure 13: Trajectories and general distribution of particles transported by wind over (left) sandy surface and (right) stony surface. Note the much higher paths of particles on the stony surface, which result from impact with larger particles.

Within all saltating curtains, the vast bulk of material travels close to the ground. Samples taken over a boulder-mantled surface show that about 50 percent of the grains travelled at levels less than 15 centimetres (six inches) and 90 percent at levels less than 63 centimetres (25 inches), even though a few grains rebounded as high as three to six metres (10–20 feet). Over cultivated fields or loose sand surfaces lower heights are recorded.

Impact creep

Particles of moderate size resting on the ground experience an abrupt forward movement if hit by a saltating grain. This is impact creep. Because a saltating grain can move particles six times its diameter and 200 times its weight by impact, this becomes an important mechanism for transport of small pebbles, one to two centimetres (0.4–0.8 inch) in diameter, which are too large to be moved directly by most winds. It also affects smaller particles, and the little jerks and spurts of grains on sand-mantled surfaces that are subject to a rain of saltating particles are of this origin.

Very fine or unusually light materials can be carried in suspension by wind when the sum of the upward-directed forces of turbulence equals or exceeds the settling velocity of the particles. Settling velocity is determined by size and density of the particle, the viscosity of air, and the force of gravity. Because most natural wind velocities are such as to produce turbulence, the problem is more one of a mechanism for getting fine particles up into the air in the first place than of keeping them there. For reasons already given, wind experiences difficulty in grasping small particles on the ground. The saltating grains unquestionably play a major role in initiating transport by suspension by blasting small particles off the ground. The first phase of most major duststorms involves saltation of grains to raise the dust that is carried away in suspension. Material that travels in suspension is most conveniently termed dust (mean diameter about 0.01 millimetre, or 0.0004 inch), and the grains that travel primarily by saltation (mean diameter about 0.25 millimetre, or 0.01 inch) may be designated as sand.

Movement of large particles

Surprisingly large rock fragments are occasionally transported by wind in unusual ways and circumstances. If material composing the ground surface is loose to a depth of an inch or two, surface stones can become perched on low pedestals owing to wind scour around their base. In time, these stones fall from the pedestals, usually in a downwind direction. They may roll or slide many centimetres in the process. In particularly loose material and with underloaded wind, scouring on the upwind side can be so excessive that a perched stone becomes tilted upwind. Eventually, it may slide off in that direction, and repeated perching and sliding can result in measurable upwind movement of such a stone. This is an anomalous behaviour and a further demonstration of the eccentricity of wind transport.

Movement of unusually large fragments, some weighing hundreds of kilograms, by wind occasionally occurs in the instance of stones resting on the surface of normally dry lake flats (playas) in arid regions. When the surface of such flats becomes wet and muddy, high winds blowing under just the right conditions can cause the stones to slide across the slippery surface leaving distinct tracks. During the winter of 1968–69, a stone weighing 6.4 kilograms (14 pounds) moved 64 metres (210 feet) across the surface of Racetrack Playa in Death Valley, California, in this fashion.

Distance of transport. The distances materials are moved by wind depend largely upon the mode of transport, terrain configuration, and wind regime and pattern. Large fragments shifted by undermining move centimetres

to perhaps a few metres at most. Particles travelling by traction, including impact creep, can cover hundreds to possibly thousands of metres under favourable conditions. Particles transported by saltation can travel tens to hundreds of kilometres, provided they are not constricted by terrain or captured by dunes, vegetation, or water bodies. There is probably an upper limit to the distance a grain can travel by saltation before it becomes completely reduced to smaller pieces by impact. This distance is not actually known but may be on the order of 160 to 325 kilometres (100–200 miles).

Inter-continental dust distribution

Materials carried in suspension cover thousands of kilometres. Dust derived from North Africa falls in England and northern Germany, 3,200 kilometres (2,000 miles) away. Ships a thousand or more kilometres at sea experience dust falls on their decks. Australian dust travels 3,000–4,000 kilometres (2,500 miles) to New Zealand. Dust thrown high into the atmosphere by volcanic explosions can circle the globe several times. Circumglobal transport of very fine particles is probably more common than realized. Very fine dust is distributed throughout the entire atmosphere. The particles provide nuclei upon which snowflakes crystallize and raindrops condense, and they are eventually removed from the atmosphere by this means.

The amount of material carried by different modes of wind transport varies with local situations, but some generalizations are possible. Traction and impact creep account for 20 to 25 percent of all sand transport; saltation provides the remainder. The mean amount of debris moved in suspension is perhaps in the neighbourhood of 20–25 weight percent of the wind-borne total. It varies widely with the local situation, and being highly visible, it is usually overestimated. Even so, individual storms have transported as much as an estimated 100,000,000 tons of dust.

EFFECTS OF WIND TRANSPORT

Modification of particles. Materials subjected to wind transport undergo modification. Wind is an excellent sorting agent, and deposits of windblown material are inevitably well-sorted; that is, the grains are of closely similar size. It is an unformulated law of nature that the sorting power of a transporting medium is the inverse of its viscosity. Glacier ice has very high viscosity, and its deposits are very poorly sorted. Air, having very low viscosity, is an excellent sorting agent.

Wind blowing over an alluvial deposit consisting of gravel, sand, and dust produces an initial sorting by removing small particles and leaving the large ones. Of the particles removed, sand is carried moderate distances and dust is carried greater distances, thus separating (sorting) these materials. The sands are further sorted by the differences in mode of transport, traction or saltation, with traction debris being the coarser. Within traction debris, larger particles travel more slowly and cover shorter distances, resulting in further sorting.

The sharp corners and edges of angular sand grains partaking of saltation are knocked off and worn away by the constant impacting. In time, these grains become rounded and smooth. Several tens of kilometres of travel by saltation is required, at least in some settings, to produce well-rounded grains. Most accumulations of dune sand, however, consist of well-rounded grains, and this is one means of identifying ancient deposits of windblown sand preserved in the geological column. By contrast, particles in deposits of dust (loess) are mostly angular and show minimal wear. This is so because they are so small and light that they are cushioned from impacts by the air itself. Glass bottles left on the ground in areas of saltating sand quickly become frosted because of little pits made by the impacting particles. Grains of vitreous (glassy) minerals, such as quartz, become frosted for the same reason.

Thus, deposits of sand grains transported by wind are distinguished by being unusually well-sorted, well-rounded, and frosted. Investigations have shown that another type of frosting can be caused by chemical processes, so frosting alone is not proof of wind transport. Investigators working at huge magnifications under electron microscopes, however, feel that they can distinguish the type of pits and cracks made by impact during wind transport from frosting produced by other means.

Deflation of surfaces. Removal of material by wind is deflation. If the mantle of loose rock debris underlying a deflating surface is a mixture of particle sizes, removal of the fine material results in residual accumulation of larger fragments. Eventually, a surface armour of coarse fragments results, and further removal of fine particles is prevented. Such residual concentrations are known as lag gravels or, in arid regions, as desert pavement. Desert pavements are smooth, firm, compact, relatively impervious, and without vegetation.

Production of desert pavement

In desert pavements the fragments, particularly if platy, are fitted together in a crude mosaic nearly completely covering the ground surface. The topsides of these stones commonly develop a dark brown to black weathering coat, rich in iron and manganese, known as desert varnish. The bottoms of these same stones, where seated in the underlying soil, usually have a brilliant orange coating rich in iron oxide.

Desert pavements form mostly on gently sloping alluvial surfaces. The pebbles are usually seated in a fine gray silt that more often than not overlies a well-developed desert soil. Desert pavements are resistant to erosion and relatively long-lived. Where well developed, they indicate an extended period of stability, with respect to processes of deposition and erosion. It is not yet a matter of full agreement that desert pavements are solely the product of wind action, some investigators preferring to believe that the stones are concentrated on the surface by a heaving process and subsurface movement of fine materials. In the classical view, however, wind is regarded as a significant genetic agent in the development of many desert pavements.

In places where the material subject to deflation is wholly fine grained no lag gravel develops, and large quantities of material can be removed. As much as 15–30 centimetres (6–12 inches) of soil have been deflated in a single season. Spots where removal is inhibited by cementation, vegetation, or by other obstructions develop positive relief; and areas where exceptional removal occurs become depressions.

Closed depressions called blowouts are generally considered to be a product of localized deflation. They can be small and shallow, involving areas of only a few square metres and depths measured in centimetres or metres, or they can be huge. Big Hollow, west of Laramie, Wyoming, is over 14 kilometres (nine miles) long, 45 metres (150 feet) deep, and nearly 10,000,000,000 tons of sand and dust had to be removed to create it. Deflation is a favoured mode of origin.

Most oases of the Sahara lie on the floors of great shallow depressions, the largest of which is 80 kilometres (50 miles) long by 30 kilometres (20 miles) wide, with depths of 100–160 metres (330–525 feet). The floors of many lie below sea level, and a number have large sand sheets and dune fields on their downwind sides. Although deflation has probably not been the only factor or process involved in shaping these depressions, they are generally regarded as largely the product of wind excavation. In Mongolia, great, shallow, closed depressions with gently sloping sides, known as P'ang Kiang Hollows, are also attributed to deflation.

Satellite photographs covering wide areas of the North African and Arabian deserts indicate the presence of extensive and prevailing topographic patterns, strongly suggesting that scour by wind plays a major role in determining the configuration of these regions. Surface travellers in these areas have long been aware of the local manifestations of wind erosion. Although wind deposition has been recognized as a major process in this area, it may be that the total effect of widespread, pervasive wind erosion has been underestimated, even though deposition at any site implies a corresponding erosion elsewhere.

Abrasion by sandblasting. Artificial sandblasting has long been employed as an efficient means of cleaning dirty buildings or etching decorative concrete, and objects impacted by natural saltating sand grains suffer similar

abrasion. Power and telephone companies operating in areas of strong eolian saltation know that the lower parts of wooden poles must be protected to keep them from being cut away by wind-driven sand. The wires of a low-strung Trans-Caspian phone line are reported to have lost half their diameter within a decade. Other artificial articles such as bricks, bottles, cans, or pieces of wood left in such environments show strong modification by sandblasting within just a few years.

The larger, harder, more angular, and faster travelling the saltation grains, the more effective is the sandblasting. It is also favoured by a ready but not too copious supply of sand, by strong prevailing winds, extended expanses of gentle terrain, and sparse vegetation. Too much sand results in accumulations that bury the objects that might be abraded.

The maximum rate of wear by sandblasting in natural situations occurs at a discrete height above ground level. This probably reflects the greater impact energy of saltating grains travelling at that height. The shape of most sandblasted faces and specific field experiments suggest that maximum wear occurs mostly between five and 25 centimetres (two and 10 inches) above the ground. At one site, subject to frequent sand saltation over a bouldery surface, maximum wear over an 11-year interval occurred at a height of 22 centimetres (nine inches).

By courtesy of Robert P. Sharp

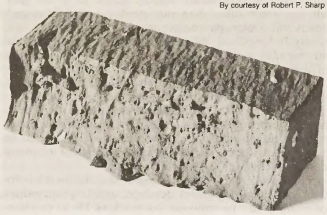


Figure 14: Common red brick after six years of natural sandblasting.

Ventifacts, flutes, and grooves

One distinctive geological result of sandblasting is the modification it produces in the shape and surface character of stones resting on the ground. These are given the special name ventifacts. Extended sandblasting can produce one or more relatively smooth, planed facets on such stones, and multiple facets intersecting in sharp edges are usual. Ventifacts are thus described as ridge-shaped (two facets), pyramidal (three or more facets coming to a peak), or irregularly polygonal (many facets). One special type of ventifact has three wind-cut faces that comprise the entire surface of the stone. This is a Brazil-nut shape and goes by the German term dreikanter ("three edges").

Not all facets are plane, some are concave or convex, especially on larger stones. On stones smaller than 15–25 centimetres (6–10 inches) facets often are cut without great regard for the stone's original shape, but this is less true of larger stones. Sandblasted faces display distinctive lustre, some are polished, others look like cellophane, and a few are greasy or dull. Additionally, they are usually pitted, fluted, or grooved in a highly characteristic manner. Surfaces that faced into the wind at angles exceeding 55°–60° are commonly pitted. Pits range from a millimetre to two or three centimetres (0.05 to about one inch) in width and depth, depending upon grain size and structure of the rock. Pits form principally in soft particles or minerals, and the associated projecting points are capped by hard grains, such as quartz or garnet. Flutes are scoop-shaped depressions, elongated parallel to the wind and of a size range similar to that of pits. Shallow, barely perceptible flutes 1–2 millimetres (0.05–0.1 inch) wide and 1–2 centimetres (0.5–1 inch) long are highly characteristic of wind-cut surfaces inclined at less than 55° to the wind.

On still more gently inclined surfaces grooves are formed. They differ from flutes in being open at both ends and commonly have dimensions approaching two or three centimetres in width and depth and lengths of many centimetres. Both grooves and flutes develop independently of structural layering in the stone, unless such structure parallels the wind. Then etching of softer layers occurs. The pattern of grooves and flutes on the upwind faces of sandblasted boulders a few metres feet in diameter shows graphically how stones of that size affect the direction of wind currents. These markings display a hemiradial pattern like shoots of water within a fire-hose stream directed against the boulder's face.

Sandblasting occurs without regard to rock type. Most well shaped, smooth surfaced ventifacts, however, are formed in relatively hard, fine-grained, homogeneous rocks, even those composed of pure quartz. Soft rocks such as limestone or marble are easily cut, but they do not long preserve the evidences of sandblasting because of subsequent weathering by solution. Friable rocks disintegrate under the impact of saltating grains and do not develop good wind-cut surfaces.

In areas of unidirectional wind, large stones show cutting only on one side, but stones smaller than 25 centimetres (10 inches) in diameter commonly display cutting on several sides. These smaller stones have clearly shifted position and changed orientation. This can happen through one of several possible mechanisms, including, among others, perching and falling from pedestals. In areas with strong winds from several directions, multiple facets should be formed without the shifting of stones, but such situations are relatively rare. Differences in the age of facets can often be recognized because of deterioration of the older wind-cut faces through weathering.

Estimates of the time required to produce a well-shaped facet on hard stones by natural sandblasting range from a decade to centuries. Laboratory tests and field observations suggest that even in areas of intense sandblasting the time is probably more on the order of a century than of a decade. Evidence of sandblasting in the form of shallow etching can be produced in a year or two on stones composed of a variety of minerals, however.

Conditions favourable to sandblasting

Sandblasting currently occurs mostly in dry, barren areas and in regions peripheral to existing glaciers, as the dry valleys of Antarctica or the margins of the Greenland Ice Sheet. Most of the world's deserts display ventifacts in local abundance. During the Great Ice Age (the last 2,500,000 to 10,000 years ago), when continental ice sheets pushed south into central Europe and central North America, conditions in areas around their margins were nearly ideal for sandblasting. Adequate quantities of loose, angular sand grains were available, strong winds blew outward from the ice sheets, and vegetation was scanty. Consequently, "fossil" ventifacts are abundant in spots throughout central Europe and North America where wind cutting is unknown today. The same is true for areas peripheral to former mountain glaciers in the Alps, Himalayas, and other glaciated ranges. Wind has generally been given more than its just share of credit for creating niches, hollows, alcoves, pits, and fretted rock surfaces as well as natural bridges, windows, and pedestal rocks. These forms are actually more the product of differential weathering, with wind helping mostly by removal of rock particles that have been loosened by weathering.

Wind scouring of modestly coherent soily materials, however, has locally created parallel U-shaped chutes separated by narrow ridges. These features, elongated in the direction of the wind and usually a metre or two wide and deep and a few tens of metres long, are named yardangs. Similar forms are developed by wind scouring, accompanied by some deposition, in the hard-packed snow that mantles large glaciers such as the ice sheets of Greenland and Antarctica. These ridges and furrows, called sastrugi, are the bane of travellers traversing such surfaces.

DEPOSITION BY WIND

Geological materials transported by wind must eventually come to rest, and eolian handling frequently results in accumulation rather than dispersal. In the immediate past, windblown dust (loess) has accumulated to thicknesses of a good many metres (10–30 feet) in numerous regions and locally to thicknesses as great as 100 metres (330 feet) in

some. Mongolia, central Europe (especially Hungary, Belarus, and Ukraine), Argentina, Uruguay, New Zealand, central Alaska, and central North America are examples of regions with extensive loess cover.

Loess forms a blanket of even-grained, homogeneous, unbedded material on the land surface. It is typically coherent enough to stand in vertical bluffs yet loose enough to cultivate easily. In Mongolia, people live in caves hollowed out of loess bluffs. Sources of dust for such deposits include desert areas such as North Africa, dust bowls such as that of the southwestern United States, raw detrital plains peripheral to glaciers, and wide river bottoms subjected to frequent flooding. Once deposited, the dust is not easily picked up again, and eventually it is fixed in place by moisture and vegetation, mostly grass.

Sand dune formation and migration. Because wind moves more sand than dust, accumulations of windblown sand are more widespread and voluminous. They occur in the lee of bushes, rocks, hills, and cliffs, in major mountain wind shadows, and in places where opposing wind regimes cancel each other. The most obvious products are the hillocks of sand known as dunes.

Growth and structure of dunes

If a heavily laden curtain of saltating sand traveling across a stony surface encounters a sandy patch, it loses velocity and deposits some of its load because of the drag effected by the loose sand surface. Thus, sand begets sand, and the patch grows in area and height. In the early stages, a gently rounded, nearly symmetrical, streamlined mound develops, and over this the wind moves smoothly without separation of the boundary layer.

As the mound gets higher, it develops a slight asymmetry with a gentler windward slope and a somewhat steeper leeward slope. Under unusually strong winds, separation of the boundary layer occurs at the dune crest, resulting in exceptional deposition of traction and saltation sand just downwind. This creates a steep lee face inclined at the angle of repose for loose sand, about 30°; at greater angles material will slide. The mound now becomes an efficient sand trap; it is truly a dune. The more sand it traps, the higher the lee face grows and the more efficient is the trapping mechanism. Sand traveling by traction and creep is dumped over the brink of the lee face, and saltating grains rain down upon it from the separated boundary layer (Figure 15). Deposition is greater on the upper reach of the lee face, causing it to become increasingly steeper. Under such conditions the slope steepens to an inclination of about 34° before it gives way by avalanching to establish a more normal angle of repose (30°–33°). For this reason, the lee slope has been called the slip face. The lee slope of an active dune is constantly being scarred by little sand avalanches all along its face. Similar avalanches can be initiated artificially by a person walking along the brink of a freshly built lee slope. The flowing sand generates a very low-pitched noise, often heard in dune areas when strong winds are blowing and active deposition is under way.

The sand composing dunes exists in two recognized and distinctive physical states well known to experienced dune travelers as accretion sand and avalanche sand. Layers of sand that accumulate under a saltating curtain, as on the windward side of a dune, become firm through the beating of the passing saltating grains. Avalanche sand of the lee slope is, by contrast, loose and poorly packed. The larger part of most migrating dunes consists of avalanche sand covered by a veneer of accretion sand. In places where the accretion is missing or unusually thin, travelers sink deeply into the loose avalanche sand.

Migrating dunes advance by deposition of material on the lee face. In time, therefore, under ideal circumstances

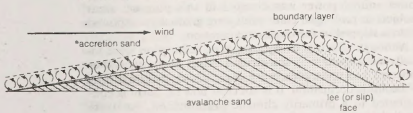


Figure 15: Distribution of sand on and within a dune due to wind action. Sand is shown raining down onto the lee or slip face; this general process leads to dune migration.



Figure 16: Sand avalanche tongues on lee face of a recently active dune.
By courtesy of Robert P. Sharp

they must consist primarily of layers of sand laid down near the angle of repose (30°–33°) inclined in the direction of dune advance. In many dunes the internal bedded structure departs from this simple model, but it has proved a useful concept in determining paleo-wind directions indicated by ancient deposits of dune sand.

Surfaces of loose sand across which wind is transporting sand usually develop a rippled configuration. The ripples are small, essentially parallel, asymmetrical ridges composed largely of sand moving by traction and impact creep, although some exchange with the saltating curtain also occurs. The windward slopes of the ridges are gentle (8°–10°), and the leeward slopes are steeper (20°–30°). At any single site, the ripples are regularly spaced at distances commonly ranging between seven and 18 centimetres (three and seven inches). The distance from crest to crest is usually referred to as wavelength, and the difference in elevation from crest to trough is height or amplitude. The height is usually less than a centimetre (0.4 inch). In cross section, most wind ripples are seen to be an accumulation of relatively coarse sand resting on a smooth surface composed of finer sand. The size of ripples—that is, their wavelength and height—is influenced by the coarseness of the grains and wind velocity, increasing with both. In some situations, giant ripples, composed largely of small pebbles, are built by impact creep. They may have wavelengths of several metres and heights of many centimetres.

Wind ripples

By courtesy of Robert P. Sharp

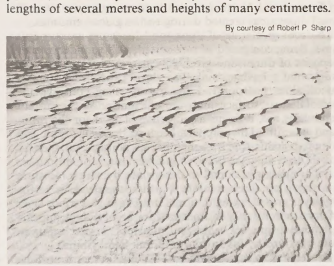


Figure 17: Wind ripples formed in sand (foreground) and in sand and fine gravel (midground).

In strong winds, it is easy to measure the rate of sand-ripple movement from reference points established by sticking small twigs into the ground and timing ripples as they move by. They move as rapidly as five to seven centimetres (two to three inches) per minute. At very high wind velocities, around 80 kilometres (50 miles) per hour, ripples disappear and the sand surface becomes smooth.

Role of vegetation. Windblown sand and vegetation display some significant interrelationships. Vegetation of almost any type inhibits transport by traction, creep, and saltation. The resulting deposition can overwhelm and kill the vegetation, although some vegetation can survive

partial or even complete burial by a migrating dune and continue to flourish after it has passed. This is likely to be particularly true of sand-tolerant plants. Some plants, known as sand-loving, thrive on active sand deposition; indeed, it seems to be required for their continued growth. These are primarily grasses with complex crawling root systems and nodes on stems from which new roots sprout in case of burial. The marram grass of Europe and North American dune grass are common examples.

Deposition of coastal dunes

Along some seashores, sand-loving vegetation captures sand blown inland from the beach and forms a long narrow dune parallel to the shoreline, known as a foredune ridge. If the beach is receiving additional sand from longshore currents (those parallel to the shore) so that it is building out into the sea, the result is a succession of foredune ridges parallel to the shore. The newest foredune ridge deprives its predecessors from the continuous supply of new sand required by sand-loving plants. Consequently, they do not thrive and are replaced by plants that are only sand-tolerant, which eventually cover the entire dune and stabilize it. Finally, these, too, are replaced by more advanced vegetative complexes, including trees. Successions of such stabilized foredune ridges occur along the North Sea and Baltic coasts of several European countries.

Not all wind deposition of sand along shorelines occurs as foredunes. Where sand supply is copious and vegetation is not able to keep pace, great sheets of sand dunes are built for miles inland. If the sand supply is later reduced or cut off, dunes commonly become overgrown and stabilized by vegetation. This sequence has been repeated more than once along some coasts because of the rise and fall of sea level related to water withdrawal during glacial periods and its return during interglacial periods. Such sea-level shifts cause major changes in shoreline position, especially on gently shelving sea floors, thereby affecting the sand supply.

The supply of sand for localized inland dune masses is often cut off for a variety of reasons involving geologic accidents or climatic changes. When this happens in humid areas, the dunes become overgrown by vegetation, rapidly assuming a "fossil" status. Subsequently, the vegetative cover may deteriorate because of climatic changes, lowering of groundwater levels, or other causes. When this happens, the sand becomes reactivated, creating local clusters of new dunes downwind from blowouts. The middle parts of Europe and North America harbour examples of such fossil dune areas created during earlier glacial episodes.

The ecological role of dune sand in desert areas merits note. Being composed of well-sorted, well-rounded grains, deposits of dune sand are highly pervious. They soak up water like a sponge, and even during the heaviest rains runoff does not occur. Furthermore, they conserve the percolated subsurface moisture efficiently. As the surface sand dries out, it forms a protective blanket for the moist sand beneath. This happens because the well-sorted and rounded grains touch each other at points, and there is much dead airspace between grains. Thus, no continuous capillary passages extend to the surface by which moisture can move up and evaporate. Further, the dead air in the dry layer provides good thermal insulation. A surface layer of dry sand only 10 centimetres (four inches) thick appears adequate to prevent significant moisture loss according to data obtained by investigation in mainland China.

Consequently, moist sand can usually be found in dune areas at depths of a few to 50 centimetres (20 inches) many months after precipitation has occurred and when the surrounding desert floor is bone dry to depths of many metres. Animals and plants are fully aware of this, and the fauna and flora associated with some desert dune masses is more varied and rich than in the surroundings. Little burrowing animals and insects live in temperature-controlled comfort. Predators come to the dunes in search of such animals, and, judging from artifacts, aboriginal people also found the dunes attractive. (R.P.Sp.)

Physiographic effects of man

Man accomplishes some physiographic changes directly—for instance, by quarrying rock or digging canals—and also contributes to the development of the Earth's surface features by influencing the performance of other agencies, as when he promotes soil erosion by clearing vegetation or initiates channel incision by building a dam. Such indirect effects are often unforeseen and may be far more powerful and long-lived than the force by which they were set in motion or directed.

In many cases, no clear distinction can be drawn between artificial and natural landforms; indeed, the direct physiographic effects of man have invited comparison with the work of glaciers, volcanoes, and termites. Though not wholly felicitous, these parallels are collectively instructive in bringing out the variety of man's handiwork. In terms of scale, the features he constructs or excavates are dwarfed by comparable landforms fashioned by other agencies; and, when viewed against the geologic time scale, the cumulative effect of man's activities appears insignificant. On the other hand, the indirect agency of man is outstanding in its pervasiveness, no physiographic process being wholly immune from its influence.

The role of man in fashioning the landscape has attracted interest since antiquity. Detailed investigation began in the late 18th century and advanced concurrently with the marked acceleration in the growth of man's power to intervene. In the course of these two centuries, man has come to dominate the physical environment of restricted areas and to affect the general environment to some degree over the entire globe. The combined pressure of technological advance and population increase has tended to enhance and accelerate the influence of man through time.

RISE OF HUMAN AGENCY

Importance of fire

Man became an important influence on physiographic change with the help of fire. Repeated burning of the vegetation cover alters its composition and hence its contribution to soil development; it can also lead to erosion by wind and water. To judge from the archaeological record, man's control of fire spans an estimated 300,000 years in Europe and Asia, more than 50,000 years in southern Africa, and at least 12,000 years in North America. The firing of vegetation in driving game and improving the yield of edible plants has been practiced by hunting and collecting communities in many parts of the world and presumably played a part in Paleolithic economics. Its cumulative effects can only be conjectured. Nevertheless, there is botanical evidence to suggest that the world's principal grasslands and prairies, as well as some major woodlands (among them the pinewoods of the southeastern United States and the teak forests of Myanmar [formerly Burma]), are the product of repeated burning over millennia. Lightning fires are too infrequent to furnish an adequate explanation. Should the agency of man be confirmed, the soils associated with the tropical and temperate grasslands would rank among the most prominent incidental effects of human agency now in evidence.

The transition from hunting and gathering to stock-raising and cultivation was slow and piecemeal; extensive territories long remained subject to temporary exploitation. The outcome was mainly to concentrate and intensify the effects of man and the animals he needed on the vegetation cover: permanent clearings spread, grazing was increasingly controlled, and preferred plant species were introduced. The trend was accentuated by the rise of farming settlements such as those to be found in many parts of Southwest Asia by 6000 BC, for they made heavy demands on timber and fuel as well as the products of plant and animal husbandry. The introduction of tillage practices enabled man to modify soil structure directly; once animal power was diverted to this purpose, areas subject to prolonged cultivation were gradually sculptured into a stepped or furrowed configuration.

Although clearing, overgrazing, and cultivation were locally accompanied by soil erosion, many of the areas thus affected had undergone much more severe erosion during Paleolithic occupation. It is not clear how far burning contributed to a primarily climatic phenomenon, but there is little justification for assuming that, until the advent of farming, man invariably respected the equilibrium between soil-forming and erosive processes.

Ancient irrigation works

Where irrigation was called for and favoured by social and physical conditions, the manipulation of soil and water left a clear imprint on the landscape. Cultivation terraces, some of them more than 3,000 years old, occupy thousands of square kilometres in Asia, Europe, Africa, and the Andes area of South America. The dikes, canals, and embankments on which the earliest urban societies depended can be traced on the great alluvial plains of the Old World. Ridged fields created in pre-Columbian times to exploit land subject to seasonal flooding have recently been identified in many parts of tropical South America. In North Africa the control of ephemeral floods in the coastal uplands awaited Carthaginian settlement and reached its apogee in Roman times; elsewhere, it was not until the Middle Ages or later that the action of running water was significantly modifed by engineering works.

Such devices had implications for landscape development that went beyond their immediate purpose. Terracing reduced the sediment yield of river basins, and the soils they held developed distinctive characteristics, especially when subject to waterlogging. In parts of Mesopotamia irrigation was not matched by adequate drainage and soon led to the accumulation of noxious salts at the surface. The embanking of streams promoted silt deposition within their channels, which, in turn, rendered them unstable and increased the incidence of flooding; the Huang Ho is a case in point. In the Tafilalt oasis of Morocco, medieval stream diversion provoked the accumulation of alluvium, reaching a maximum depth of five metres (16 feet).

Where runoff had been brought under control by damming, as in Roman Libya, or by underground drainage, as in Etruscan Italy, valley erosion was checked, and the topography now contrasts sharply with that of river basins not thus protected. Conversely, irrigation made it possible to cultivate marginal areas where the role of vegetation in shielding the soil from water and wind action was especially important; the Romans, at other times hailed as supreme practitioners of conservation, have on these grounds been held responsible for the decay that ultimately overtook large parts of their empire.

Materials for the pyramids and other works

The building of cities, engineering works, harbours, monuments, and defenses had as its counterpart the extraction of large volumes of rock and clay. The Sadd-el-Kafara, a dam built south of Cairo between 2950 and 2750 BC, alone embodied 100,000 tons of material; the Great Pyramid of Khufu at Giza consists of 2,300,000 blocks with an average weight of 2.5 tons. Most of the material used in construction was obtained from quarries and floodplains.

Shafts had been employed in Neolithic flint mines throughout western Europe, notably at Spiennes, Belgium, where the resulting pitted landscape extends over two hectares (five acres). By the early 4th millennium BC, shafts were in use in the copper mines of Armenia and the Caucasus. Shafts also made possible the digging of tunnels called *qanāts* or *foggara* that were as long as 20 kilometres (12 miles) in order to tap the underground water held by alluvial deposits; the aligned craterlike depressions by which they are marked are a prominent feature of the Iranian desert, where the technique has been known for 3,000 years. The scars of mining and of other forms of excavation spread with the adoption of new minerals and fuels and the dispersion of technical knowledge; *qanāts* are to be found in other parts of Asia and in Europe, North Africa, and the Americas.

Aerial photography has played an important part in filling in the details of the ancient man-made landscape. It has revealed the grid laid out by Roman surveyors in Tunisia and other parts of the Mediterranean, the puzzling Nazca designs of the Peruvian desert, and many mounds and dikes that fulfilled military and religious functions. The picture that emerges is one of gradation between areas subject to long-standing, though varied, direct and indirect modification and others where man's handiwork betrays short-lived, opportunist use of the land.

EFFECTS OF MODERN MAN

In reviewing the effects of modern man on the landscape, it is convenient to classify human activities in terms of the geologic processes to which they most closely correspond. The issue of motives, and any assessment of the outcome in economic or aesthetic terms, is not central to the present discussion.

Direct effects. The technological and demographic upsurge that characterized the 18th and 19th centuries was marked by a rapid increase in the power of direct human intervention; by 1922 it was possible to claim that, in the densely populated British Isles, man's capacity for removing and comminuting rock exceeded that of all the atmospheric forces combined.

Land cultivation and mining. The uneven spread of this power is illustrated by the distribution of cultivation methods. About 11 percent of the world's land area is under cultivation, of which one-tenth is irrigated. The technology employed ranges from digging sticks to mechanical devices that wholly transform the existing soil structure and from the occasional diversion of spates to the application of specified volumes of water at controlled rates. Soil chemistry may be modified by the abstraction of nutrients by plants until their exhaustion, or it may be manipulated by rotation and chemical fertilization. Other weathering agents are influenced: frost action may be encouraged to improve tilth or may be checked if injurious to the crop.

The boundaries of the sown areas continue to fluctuate. In New England, large areas have reverted to forest in the course of the last century; conversely, drainage has added more than 600,000 hectares (1,500,000 acres) to the cultivable territory of Finland since 1941. The margins of arid areas are rendered especially unstable by rainfall variability.

The relief produced by plowing and by the excavation of drainage and irrigation canals is limited by the stability of the soil material as well as by the available power and equipment. In this respect it conforms in vertical scale to the products of water erosion, however much it may depart from the fluvial landscape in its geometric pattern.

Quarries and excavations

In contrast, the depths attained by open-pit mining are determined primarily by the economics of mining. The copper pit at Bingham, Utah, is more than 800 metres (2,600 feet) deep and has not yet reached the limits of rewarding excavation. Laterally, the scope is technically unlimited, although the largest quarries, such as the Hull-Rush-Mahoning iron ore pit in Minnesota, remain far smaller than the Aso caldera of Japan and other volcanic features.

Open-pit mining accounts for 70 percent of the world's output of minerals and rocks. The excavations range from small borrow pits, such as the 300,000 marl pits that dot East Anglia, England, to vast stepped amphitheatres. The strip-mining of coal produces its own distinctive topography, as does the working of placer deposits by hydraulic methods or by dredging. The terrain thus affected is rapidly expanding; in the United States, for example, the production of natural aggregates increases by 5 percent each year. Because sand, gravel, and rock suitable for crushing are low in value, their extraction tends to be concentrated near urban and industrial sites. This tendency is countered, however, by the needs of highway and aqueduct construction, and it may be largely ignored where more precious minerals are involved, as in the case of the great copper open-pit mine of Chuquicamata in the Atacama Desert of Chile. Military and scientific activities, such as the explosion of atomic devices and the creation of craters by bombing, have further dispersed man's erosional activities.

Man-made pits are often bordered by waste tips and mounds. The proportion of usable to unusable solids ranges from less than 20 percent of the output in sands and gravels to as much as 14,000 percent in the case of slate. Contributing to the process are such underground workings as the gold mines of the Witwatersrand, South Africa, which yield some 60,000 tons of waste each year. The potential instability of some of these features has been borne out by the collapse of tips at Espenhain in what was East Germany in 1959 and in Aberfan, Wales, in 1966. The unconsolidated material of which they are composed has constructional value, and, together with urban and industrial waste, it is employed to level and create building land in response to economic and social pressures.

Artificial channels and coastal works. Surface waters are more widely affected. The density of the drainage network is increased by canals, ditches, and other artificial channels. Surface drains may triple the length of channel that would otherwise characterize a particular area; in parts of Scotland the increase has been 20-fold. Conversely, subsurface drainage can lead to the local elimination of organized runoff. Canals, including the Panama cut, exploit existing watercourses and also disregard watersheds; the channel that links Abidjan, Ivory Coast, with the sea bears little relationship to the natural drainage. The geometry of the channels and their floodplains is often modified to improve the efficiency of drainage; dikes, embankments, and auxiliary canals remain a prominent feature of the fertile alluvial valleys of the Old World and have transformed the valley of the Mississippi River. Lake Kariba, over 280 kilometres (170 miles) long, on the Zimbabwe–Zambia border, is the outstanding example of further diversification produced by artificial lakes and reservoirs.

Harbours, breakwaters, and dikes

Coastal modification similarly ranges from the minor adjustment of existing features to the creation of novel landforms. Most of the major harbours of the present day, including those of Sydney, San Francisco, and New York, occupy propitious sites; a few are largely artificial, such as those of Takoradi, Ghana; Sète, France; and Algiers, Algeria. The port of Ischia, Italy, lies on a crater lake that has been opened to the sea. Offshore islands are also exploited in the building of breakwaters; artificial islands—for instance, Rincon Island in the Gulf of Mexico—may be the solution to the problems posed by offshore drilling.

Protective works range from those designed to preserve the existing situation, such as sea walls, to groins and other devices that work in conjunction with the transport by waves and currents. The creation of land at the expense of the sea has been most successfully pursued in areas already subject to coastal sedimentation. In England, the area around the Wash has gained 16,000 hectares (40,000 acres) of land since the 18th century. In the Netherlands, over 740,000 hectares (1,800,000 acres) have been won in this way since the early 13th century; the dike that blocks the IJsselmeer lies at a maximum distance of 85 kilometres (53 miles) from the former coast and has reduced the length of the coastline by about 300 kilometres (200 miles). The material yielded by the dredging of existing facilities is sometimes used for reclamation rather than being dumped at sea: a combination of dredging and filling was employed in developing San Juan Bay, Puerto Rico.

Man's direct intervention is seen to differ from other physiographic processes chiefly in its versatility. It is not unique in effecting both chemical and physical changes and transcending the distinction between the organic and inorganic worlds. Tree roots, for example, break up rocks by mechanical means and also influence microbial activity. But human agency is not confined to a specified range of environments or sources of energy. Glaciers and rivers are creatures of topography and climate; volcanoes are restricted to well-defined zones; and even the action of wind is strongly localized. Man, however, is potentially free of such constraints.

Indirect effects. The incidental effects of man's activities on the landscape have long attracted attention. Plato, for instance, ascribed the barrenness of Attica to soil erosion consequent upon deforestation; Pausanias pointed out that silting at the mouths of the Achelous and Maeander rivers was accelerated by cultivation within their catchments. Later writers came to regard such phenomena as indicative of the tendency by modern man to disturb the equilibrium attained by nature, a viewpoint that was powerfully endorsed by the soil erosion that afflicted the U.S. Midwest in the 1930s.

The difficulties to be faced in measuring such unforeseen and often undesirable developments have long been emphasized. The need remains to rely on historical evidence and on deduction based on present-day parallels, and many of the effects are still poorly understood. Nonetheless, advances in the study of soil mechanics, meteorology, geophysics, and allied fields have thrown light on some of the mechanisms by which indirect human agency operates and have demonstrated that the consequences are sometimes out of all proportion to the energy involved in man's intervention.

Land subsidence and slope failure. In many cases the role of man is self-evident and easily evaluated. The underground extraction of solids and fluids, for example, can lead to subsidence, and its prediction presents few problems. The salt fields of Cheshire, England, are widely affected by such activity, and flooding of the depressions has formed lakes known locally as meres or flashes. The coal-mining areas of Pennsylvania provide instances both of subsidence and of collapse. Where water, oil, or gas is withdrawn from an underground stratum that is poorly consolidated, the loss of supporting pressure provided by the fluid can result in compaction. Mexico City has been subject to subsidence since pre-Columbian times because of drainage by surface channels; the process was greatly accelerated by pumping to supply urban needs, and sinking at the rate of 15 centimetres (six inches) per year was observed in places. Despite restrictions placed on pumping, parts of the city are still settling by about three centimetres each year. In the Wilmington, California, oil field, the abstraction of petroleum and gas has produced depression of the surface over an area of about 56 square kilometres (22 square miles), to a maximum depth of nine metres (30 feet). The application of large surface loads also leads to depression; at the site occupied by Lake Mead, Nevada, for example, a maximum of 175 millimetres (seven inches) of subsidence has occurred in the course of 15 years, a figure that comes close to the predicted elastic yield of the Earth's crust.

Underground fluids

Analogous topographic changes can arise through drainage or irrigation of the soil. The drainage of organic soils promotes their shrinkage through the decomposition and oxidation of organic matter. Where the superficial deposits consist of loose, dry material of low density, such as windlain particles, the application of water sometimes produces the phenomenon of hydrocompaction. Over 500 square kilometres (200 square miles) have been thus affected in the San Joaquin Valley of California, and in places depression amounts to more than five metres.

Slope failure, like subsidence, occurs when gravitational stresses overcome the resistance of the rock material (see above *Slope movements*). Man may bring this about by undercutting and steepening slopes, or by changing the moisture content and hence the cohesion of deposits in which they are developed. The former is illustrated by the extensive landslides that took place during construction of the Panama Canal, one of which involved 400,000 cubic metres (14,000,000 cubic feet) of material. Infiltration from gardens and cesspools allegedly triggered an extensive landslide in the Palos Verdes Hills of Los Angeles in 1956; in three years movement exceeded 20 metres (70 feet). The Vaiont Dam in Italy was destroyed when a landslide comprising over 240,000,000 cubic metres (8,500,000,000 cubic feet) slid into the reservoir; by raising the water table, the dam itself had contributed to slope instability. In areas with a permanently frozen subsoil (permafrost), thawing may be encouraged by the disturbance or removal of vegetation or superficial sediments and soils, since these act as insulators. Highway and railroad construction in Alaska has been accompanied by landslides, slumps, and rock flows, and road surfaces are distorted by frost heaving. The vibration produced by passing traffic may also trigger movement in unstable slopes.

Techniques have been developed to preserve or restore slope stability, for example by creating an insulating layer over frozen ground or by controlled drainage of terrain weakened by seepage. The Folkestone and Warren landslips of Kent, England, have been controlled by drainage combined with coastal protection, because undercutting by wave action is an important contributory factor. In California, waterlogged clay slopes have been dried by circulating warm air through tunnels. The use of vegetation to bind the surface of cuttings and embankments is commonplace.

Erosion and sedimentation. Any change in the geometry, water discharge, or sediment load of a stream will disturb existing equilibrium conditions, and is followed by readjustment that may involve erosion, deposition, or

both. The silt that accumulates behind a dam often forms a delta; in the case of the Furnish Dam, Oregon, the entire basin was filled in 16 years. By trapping sediment and altering the pattern of flow, dams often lower the river bed in the valley below the dam. The channel of the Red River of Oklahoma and Texas has in this way become both wider and deeper for a distance of 160 kilometres (100 miles) below the Denison Dam. River straightening and dredging sometimes lead to further incision by accelerating the velocity of flow, as in the Rhine, which has deepened its bed at Duisburg by two metres (seven feet) since 1900. On the other hand, man may promote deposition by introducing surplus sediment, as in the Sierra Nevada of California and other areas subject to hydraulic mining.

The incidental effects of stream modification extend to lakes and coasts, whether as changes in the water balance or in terms of the sediment supplied. The progressive fall that has characterized the Aral and Caspian seas in recent years is attributed largely to dam construction on their affluents (inflowing streams); it has exposed 38,000 square kilometres (15,000 square miles) of land in the former and 25,000 square kilometres (10,000 square miles) in the latter. The area of San Francisco Bay has contracted by over 700 hectares (1,700 acres) since 1850, largely because of the import of sediment by rivers affected by hydraulic mining of placer gold; in contrast, Ventura Beach, California, has been deprived of much of its sand supply by the damming of the Ventura River.

Urbanization and hydrologic change

Urbanization is a further source of hydrologic change. By paving the Earth, man armours it against weathering and erosion; he also renders it impermeable and thus produces a far higher proportion of runoff (as opposed to infiltration) than would otherwise prevail. A little less than 1 percent of the world is occupied by buildings, roads, and kindred structures, but their distribution is uneven. Measurements taken in the United States have shown that, during construction, building sites release sediment at rates up to 40,000 times greater than those of corresponding rural areas under grass or woodland; for urbanized and developing zones as a whole, the factor ranges between 10 and 100. Channel flow gains little from groundwater and is therefore closely related to the incidence of precipitation; runoff is rapid, and the proportion lost to sewers is quickly restored to surface drainage. Hence, discharges are relatively higher and more short-lived than in areas where infiltration plays a prominent role in the hydrologic cycle and, in turn, contribute to stream load by producing channel enlargement. In brief, the progress of urbanization enhances stream activity, always providing that some erodible material has been left exposed to the action of water.

Like river basins, stretches of shoreline operate as physiographic units that adjust to man-made modification by changes in geometry or in process. Where breakwaters, jetties, or groins are erected to trap material, further erosion may take place along the coast if the beach depends for its equilibrium on the supply of sand or shingle by longshore movement of sediment. Much agricultural land has been lost in this way near Durban and Madras, and a breakwater erected to form a harbour at Santa Barbara, California, caused silting in the harbour and erosion along more than 15 kilometres (nine miles) of coast. The extraction of shingle can be equally deleterious by robbing the beach of its protection against wave action; it has aggravated erosion along the Holderness coast of England, indicated by the loss since Roman times of a strip averaging four kilometres (two miles) in width.

At times man intervenes to restrain erosion, as when he fixes coastal dunes with the help of vegetation, or seeks to establish or restore a balance between coastal erosion and deposition. In England, trucks have been employed to carry shingle from the beaches of Rye to those from which it had been derived at Winchelsea, and a similar function is fulfilled by pumping sand across the erosive channel that leads to Palm Beach, Florida. To maintain or create a deep navigable channel, however, submarine erosive processes are turned to advantage by means of training walls or by judicious dredging. Liverpool Docks and the New York Harbor both benefit from this stratagem.

Soil erosion

Man's indirect influence is displayed most prominently in the varied physiographic changes that are grouped under the heading of soil erosion and that affect two-thirds of the Earth's land surface. In the United States, for example, 5 percent of the land suitable for cultivation or forestry has been entirely ruined, 15 percent has lost three-quarters of its topsoil, and 41 percent is subject to moderate erosion. In addition to its local implications, soil erosion affects other landscape elements and the performance of allied processes. The incidence of violence of flooding is increased by rapid runoff and low infiltration; stream channels are choked by silt; and material removed by wind erosion accumulates in drifts.

Many factors help to determine the stability of soils, and the scope for human intervention is correspondingly large. Differences in the character and density of the vegetation cover generally account in large measure for areal variations in the severity of soil erosion; hence, the importance of fire and burning, as previously discussed. In the United States it has been found that land used for row cropping may erode at rates that are 100 times greater than for woodland or pasture; in Cyprus, the sediment yields from vineyards are more than six times those on comparable terrain under forest cover. In either case, slight modification of the vegetation cover will produce disproportionate changes in the rate of soil loss.

Soil structure provides further opportunities for inadvertent triggering or acceleration of erosional activity. Vegetation and plant litter protect the soil against the impact of raindrops and the action of running water and wind. A soil structure that is resistant to these forces may be destroyed by compaction, whether by hooves or vehicle tires. In many badly eroded areas, gullies are fed by runoff from roads that have not been provided with adequate gutters. Once linear erosion is initiated, soil moisture is locally depleted; the benefits of good soil structure are overridden by downcutting and bank collapse.

Unlike vegetative cover and soil structure, whose reinstatement is hampered and ultimately precluded by erosion, the factor of topography is amenable to correction; hence the emphasis on terracing and contour plowing in many soil conservation schemes.

Climate and air pollution. Rainfall intensity and the duration of individual storms are important factors influencing flood flow and the protective value of the vegetative mat, particularly in semiarid areas. Although large-scale weather modification by man is still in its exploratory stages, there is little doubt about man's capacity to stimulate precipitation by seeding suitable clouds with condensation nuclei. Hail suppression by this means is already practiced, and the mitigation of destructive downpours appears feasible. The climatic changes associated with urbanization are not wholly irrelevant to erosion, because they may involve increases in the total annual precipitation of up to 18 percent, a value that will be enhanced by the impermeability of the terrain on which it falls.

Air pollution, which is an important element of urban climates, may be influential in rural areas if it is virulent enough to weaken the vegetation cover. The emission of sulfur dioxide by the copper smelters of Ducktown, Tennessee, has had this effect and thereby led to severe gullying; a similar chain of events has been reconstructed for the Swansea Valley of Wales in the 18th and 19th centuries. It is also suggested that overgrazing in parts of the Rājasthān desert of India has increased the local dust cover and that this, in turn, leads to desert formation.

Air pollution is suspected of influencing world climate as a whole. The carbon dioxide content of the atmosphere has increased by an estimated 11 percent over its value in 1850, and the associated warming (greenhouse effect) is put at 0.6 percent between 1880 and 1940. It has long been held that the combustion of fossil fuels is a primary source of the additional carbon dioxide; the oxidation of humus, peat, and other organic material is also regarded as a major contributor, but even here man retains some responsibility because the process is encouraged by forest clearing, cultivation, and the drainage of bogs. Overall, the extent of human influence on long-term temperature changes has not yet been precisely determined.

Earthquakes. Forces acting within the Earth are not immune from the influence of man. It has been observed that the incidence of earthquake shocks has increased near the sites of large reservoirs, among them lakes Mead, U.S.; Kariba, Zimbabwe; Koyna, India; and the Arapuni River, New Zealand. On the other hand, there is some evidence to suggest that Lake Mead has reduced earthquake activity, as the seasonal variations in its level during 1939–51 show some correlation with the number of shocks within a given radius. Underground nuclear explosions are followed by aftershocks, and some believe that they may trigger impending earthquakes in the vicinity. This has given rise to proposals for the controlled release of strains within the crust by means of explosions, in order to avert major earthquakes. The finding that the leakage of fluids into fault zones appears to affect their activity has similarly encouraged attempts to lubricate lines of crustal fracture subject to intermittent but large-scale displacement.

Interaction of human and natural forces. Other environmental changes that have been attributed wholly or in part to human agency can be explained in other ways. Many of the erosional features that are regarded as characteristic of accelerated erosion are duplicated in the geological record for periods apparently immune from human depredation. The sediment yields and erosional rhythms they represent are in some cases found greatly to exceed those now prevalent. In New Mexico, for example, ancient arroyos appear to have formed at a rate that is more than nine times greater than those of the immediate prehistoric and historic period; and in some valleys in Iowa the deposition of alluvium in prehistoric times took place at over twice the modern rate. These findings challenge the view that a placid physiographic regime was disrupted by abuse of the land, although the possibility that abuse of the land is a long-standing attribute of man has already been raised. An alternative approach is to postulate changes in the climatic conditions governing physiographic processes. In the Mediterranean area, for example, erosional features often attributed to medieval deforestation and overgrazing can be interpreted in terms of a slight change in the seasonal incidence of rainfall. The onset of gullying in the southwestern United States is attributed by some to the arrival of white settlers armed with herds of cattle and metal plows, and by others to a climatic change that led to a weakening of the protective grass cover and perhaps also to a higher incidence of erosive floods.

In many such cases the answer probably lies in the combined effect of human and climatic influences. Man's potential for influencing other physiographic processes is understood; the difficulty lies in isolating his contribution to the landscape, which is a prerequisite for effective control of deleterious physiographic change. (C.V.-F./Ed.)

BIBLIOGRAPHY

Comprehensive works: R.W. FAIRBRIDGE (ed.), *The Encyclopedia of Geomorphology* (1968), a many-authored, international, one-volume encyclopaedia of geomorphic terms, processes, and studies, with precise definitions and relevant source literature; and L.C. KING, *The Morphology of the Earth,* 2nd ed. (1967), provides a detailed discussion of the Earth's surface features and the fundamental forces affecting their development.

(Ed.)

Tectonism: The following books give general but moderately technical treatment to the physiographic effects of tectonism. The *Encyclopedia of Geomorphology,* ed. by R.W. FAIRBRIDGE (1968), includes articles on all major aspects of the physiographic effects of tectonism. For a more modern but somewhat technical account of all physical attributes of the Earth that are related to tectonism, see P.J. HART (ed.), *The Earth's Crust and Mantle* (1969).

(I.P.G./Ed.)

Weathering: Weathering is discussed in some detail in several general geomorphological and geological texts. B.W. SPARKS, *Geomorphology,* ch. 3 (1960), is notable for its critical approach. A. HOLMES, *Principles of Physical Geology,* 2nd ed. (1965); and W.D. THORNBURY, *Principles of Geomorphology,* 2nd ed., ch. 4 (1968), are quite comprehensive. Several aspects of weathering and its results are illustrated in C.R. TWIDALE, *Geomorphology, with Special Reference to Australia,* ch. 4 and 5 (1968), and with M.R. FOALE, *Landforms Illustrated* (1969).

(C.R.T.)

Slope movements: LAURITS BJERRUM, *Mechanism of Progressive Failure in Slopes of Overconsolidated Plastic Clays and Clay Shales* (1966), a modern theoretical analysis of slope stability in overconsolidated clays; EDWIN B. ECKEL (ed.), *Landslides and Engineering Practice* (1958), a summary of research on the origin and correction of landslides in the United States with respect to highway construction; ALBERT HEIM, *Bergsturz und Menschenleben* (1932), a report on rock falls and mountain slides in the Alps, demonstrated by many examples; S.E. HOLLINGWORTH, J.H. TAYLOR, and G.A. KELLAWAY, *Large-Scale Superficial Structures in the Northampton Ironstone Field* (1944), the first description of slope deformations originated by bulging in valley bottoms; C.F.S. SHARPE, *Landslides and Related Phenomena* (1938), a survey of slope movements, their classification, and relationship to geomorphological cycles; KARL TERZAGHI, *Mechanics of Landslides* (1950), a classic work on processes leading to landslides and considerations on modern preventive measures; QUIDO ZARUBA and VOJTECH MENCL, *Landslides and Their Control* (Eng. trans. 1969), a book summarizing the economic importance of landslides, their relation to geological conditions, and controlling measures—includes a comprehensive list of literature.

(Q.Z./Ed.)

Fluvial processes: B.R. COLBY, "Fluvial Sediments: A Summary of Source, Transportation, Deposition, and Measurement of Sediment Discharge," *Bull. U.S. Geol. Surv.* 181-A (Nov. 1963), a brief review of the range of river sediment relationships; F.M. HENDERSON, *Open Channel Flow* (1966), a modern text in the hydraulics of flow in all kinds of open channels, includes several excellent chapters on sediment transport and the behaviour of natural rivers not found in any other text, for the advanced student; S. LELIAVSKY, *An Introduction to Fluvial Hydraulics* (1955), contains varied treatment of specific topics related to natural channels and canals, includes information from river control works in Europe; L.B. LEOPOLD, M.G. WOLMAN, and J.P. MILLER, *Fluvial Processes in Geomorphology* (1964), a textbook in analysis of landforms with primary emphasis upon rivers, contains much empirical information on rivers, as well as brief treatment of related hydraulic phenomena; H. ROUSE (ed.), *Engineering Hydraulics* (1950), a treatise containing separate chapters on a variety of topics in hydraulics ranging from hydraulic machinery to sediment transport, for the advanced student.

(M.G.W.)

Processes of glaciation: C. EMBLETON and C.A.M. KING, *Glacial and Periglacial Geomorphology* (1968), a scholarly, comprehensive treatment of the origin of glacial and periglacial landforms, written for the student with some previous knowledge of geomorphology or physical geography, and including many references to original works, organized on a chapter-by-chapter basis; R.F. FLINT, *Glacial and Quaternary Geology* (1971), encyclopaedic and world coverage of all aspects of Pleistocene glaciation, Quaternary climatic changes, the fossil record, and geomorphic history of nonglaciated regions, including a bibliography of more than 1,300 entries covering the relevant literature of the world.

(J.H.Z.)

Wind action: R.A. BAGNOLD, *The Physics of Blown Sand and Desert Dunes* (1941), the scientist's bible for the mechanics of wind action, based on extended field observations and laboratory experiments; E.E. FREE, *The Movement of Soil Material by the Wind* (1911), the most comprehensive and best documented description of wind work in the English language; W.F. HUME, *Geology of Egypt,* vol. 1, ch. 3 (1925), a comprehensive review of wind effects and action in a part of the Sahara, a classic locality; W.S. COOPER, *Coastal Sand Dunes of Oregon and Washington,* pt. 2, pp. 25–76 (1958), an excellent digest of basic principles bearing on wind, sand, dunes, and relationships with vegetation; E. IRVING, *Sand Control Research in Communist China* (1962), a translation of reports on 23 papers given at a symposium in Peking in 1960, provides excellent insight into research on sand control and sand mechanics by a large number of Chinese scholars.

(R.P.Sp.)

Physiographic effects of man: R.L. SHERLOCK, *Man As a Geological Agent* (1922), deals with the British Isles in unusual detail. Physiographic change is one of the many topics treated in an international symposium on *Man's Role in Changing the Face of the Earth,* ed. by WILLIAM L. THOMAS, JR., JR. (1956). ROBERT F. LEGGETT, *Geology and Engineering,* 2nd ed. (1962), considers many of the problems that confront the civil engineer. The use of air photographs in tracing prehistoric, classical, and medieval man-made features is illustrated in JOHN BRADFORD, *Ancient Landscapes* (1957). GEORGE PERKINS MARSH, *Man and Nature* (1864), is the first general study on the subject.

(C.V.-F./Ed.)

German Literature

German literature consists of the literary works of the German-speaking peoples of central Europe. Its development having transcended oft-changing political boundaries, it includes not only the writings from what are now the two republics of Germany but also those from Austria and Switzerland.

This article provides a concise historical survey of German literature. Its major periods and movements are discussed in relation to broader cultural developments throughout Europe and its ties with or indebtedness to other literatures are noted.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 621, and the *Index*.

The article is divided into the following sections:

Old High German period 27
Middle High German period 27
  Characteristics of the period 27
  Principal literary forms 28
    Court epics
    Courtly lyric
    Prose and drama
From Middle High German to Baroque 28
  The courtly revival and the humanist movement 28
    The Burgundian "renaissance"
    Humanism
    Austrian literature under the Habsburgs
  The Reformation 29
    Luther's influence
    Swiss-German literature of the Reformation
    Other writings of the time
  The close of the 16th century 29
The 17th century 30
  The character of the period 30
  Major works of the period 30
    Lyric poetry
    Drama
    Prose narrative
    Philosophy and criticism
The 18th century 30
  The age of Enlightenment 30
    Rationalism
    The reaction against rationalism
    The influence of Lessing
  The age of Goethe 31
    Sturm und Drang
    Neoclassicism
  The Romantic Movement 32
    First phase
    The second Romantic school
The 19th century 33
  Grillparzer, Büchner, and the drama 33
  Lyric poetry 33
    Heine
    "Young Germany"
  Realism and regionalism 34
  Naturalism 35
The 20th century 35
  Major literary trends and conditions 35
    Impressionism
    Symbolism
    Expressionism
    Post-Expressionism and Social Realism
  Literature after World War II 37
Bibliography 38

Old High German period

Although Gothic was the earliest recorded Germanic language and the only east Germanic language on which there is trustworthy information, only a translation of the Bible by Ulfilas (*c.* 311–*c.* 383), a Gothic bishop, has survived in fragments.

The first written records of the western German tribes dated from the second half of the 8th century. There is, however, evidence of an earlier, orally transmitted literature, consisting of short *Heldenlieder* (songs celebrating the exploits of famous heroes), hymns connected with pagan religious rites, battle songs, and laments for the dead. Although none of these was recorded, their substance formed the basis of later popular heroic epics. As in other areas, the first significant texts were a product of the efforts to spread Christianity. Many were translations from Latin, as, for example, the Alemannic version (*c.* 800) by an unknown author of *De fide catholica* ("Concerning the Catholic Faith") by the Spanish churchman and scholar Isidore of Seville (*c.* 560–636) and the translation of Boethius' *Consolation of Philosophy* made about 1000 by Notker Labeo of the Swiss monastery of St. Gall.

Heldenlieder

The few extant verse works were more original, albeit religious and didactic in subject. The largest, the *Evangelienbuch* (*c.* 870; "Book of the Gospels") by Otfrid of Weissenburg, presented the life of Christ in a form rivaling that of the *Heldenlieder*. The poem is noteworthy as the first German work to replace the alliteration of Germanic tradition with the end rhyme of Medieval Latin verse.

Pagan works

Despite the church's opposition to interest in anything pagan, there survived a few pagan works, such as the *Zauberspräche*—magic spells used to protect domestic animals and for various other everyday concerns. The fragmentary *Hildebrandslied* (*c.* 800; "Song of Hildebrand")—a grim account of a duel between father and son—is more interesting and is important as the sole relic of heroic verse in Old High German. Throughout the period the vernacular was sporadically used as a literary medium, but scholars continued to prefer Latin. The movement for ascetic reform, which spread from the monastery of Cluny in France all over western Europe in the first half of the 10th century, discouraged churchmen from writing for the laity, and hence there was even less recorded literature in the vernacular for about a century. By the time the vernacular was written again, important changes had occurred in the language and the old ecclesiastical didacticism had been replaced by courtly feudalism.

Middle High German period

CHARACTERISTICS OF THE PERIOD

The changes accompanying these events marked the transition from Old to Middle High German, one important point being that the knight replaced the cleric as a poet. Early works still reflected the clerical tradition, but already in Heinrich von Melk's *Von des tôdes gehugede* (*c.* 1160; "Remembrance of Death") the feudal knight's love of combat was used to illustrate the transient nature of life. The main literary forms at this time were *Minnesang* (love lyric) and epic. Poets, mainly noblemen, expressed their love according to courtly convention or told traditional tales of combat and romance.

It was long believed that the epic of local tradition was created by the *Spielmann* (wandering minstrel), who collected and recited heroic songs on popular legends and began to link together the *Heldenlieder* to form longer epic narratives. Many scholars have abandoned this view, however; the minstrels are believed to have been illiterate and poorly educated, and there is no evidence that they composed these works. Probably the earliest such composition was *König Rother* (*c.* 1160), a repetitive tale of violent bride abduction. The poets often provided interesting, sometimes amusing, commentaries on the life

of their times. *Salman und Morolf* treated the conflict between Christian and pagan communities, reflecting an interest aroused by the Crusades also evident in *Orendel* and *Sankt Oswald.* With the *Alexanderlied* (*c.* 1130), a free rendering of the *Roman d'Alexandre,* German poets began drawing on successful French epics.

The *Nibelungenlied*

The anonymous Austrian poet who wrote the *Nibelungenlied* (*c.* 1200–10; "Song of the Nibelungs") displayed great dramatic skill and poetic ability in recounting a well-integrated story of the hero Siegfried that combined songs of heroic legends about historical events and an epic concerning the destruction of the Burgundians, or Nibelungs, in 437. *Gudrun,* composed around 1210, centred on the heroine Gudrun's steadfastness in the face of abduction, while a collection, *Das Heldenbuch* ("The Book of Heroes"), includes romances about Theodoric the Great.

PRINCIPAL LITERARY FORMS

Court epics. The nationalistic epics, popular principally in the south, were counterbalanced in the west by court epics based on French models. The first notable work, by Eilhart von Oberg, dealt with the Arthurian subject of *Tristrant und Isalde* (*c.* 1170). Although Dutch, Heinrich von Veldeke (Henric van Veldeke) established himself as the father of the German court epic with his *Aeneid* (*c.* 1175–86), which was based on a French source. The three most famous poets of the court epic were Hartmann von Aue, Wolfram von Eschenbach, and Gottfried von Strassburg. Their work, dated largely between 1190 and 1210, was concerned with the knightly virtues of moderation, constancy, loyalty, and the duty of service to superiors and to God. Hartmann was an artistic and lucid storyteller. In *Erec* and *Iwein,* based on the French romances by Chrétien de Troyes, he handled the conflict between private inclination and public responsibility that existed for the medieval knight, and in *Gregorius* and *Der Arme Heinrich* ("Poor Heinrich") he illustrated man's relationship with God. Wolfram von Eschenbach's *Parzival* employed highly original imagery in a subtle and ambitious treatment of man's search for truth in his relationship with God, its hero Parzival progressing in maturity to become keeper of the Holy Grail. Gottfried von Strassburg took Thomas of Brittany as his source for his *Tristan;* a careful master of form and artistry, he was less concerned with otherworldly considerations and concentrated on an ideal primacy of love in this life.

The foremost poets of the court epic

Later epic poets could not equal these three. Heinrich von Türlin's epic *Die Krone* ("The Crown") followed Hartmann but was badly written. Despite the prolific writing of Rudolf von Ems and Konrad von Würzburg, the epic tradition gradually weakened. Although *Meier Helmbrecht* (*c.* 1250), by Wernher the Gardener, movingly described the lawlessness and violence of decadent feudal society, the social level of its subject and its realistic approach announced the end of courtly literature.

Courtly lyric. The other important literary form, *Minnesang,* the courtly love lyric, like the court epic, borrowed its content and stanza form from French or Provençal models. Its poets were minor noblemen who often became court poets to the higher nobility or traveled as minstrels throughout Germany, Austria, and Switzerland. Many of the poems representative of *Minnesangs-frühling* ("springtime of *Minnesang*") were recorded in an early 14th-century manuscript collection, the Codex Manesse, attributed to Swiss-born Rudiger Manesse (1224–1304).

The love lyrics of Walther von der Vogelweide

About 1190 Reinmar von Hagenau became court poet to the dukes of Austria, shortly after he had tutored Walther von der Vogelweide. Walther's love lyrics, which united the courtly style for which Reinmar became famous with popular, natural love poetry, were the finest produced in medieval Germany, and he influenced later poets throughout Austria and Tirol. His *Sprüche*—moral and political poems—dealt with his relationship with patrons, the struggle between empire and papacy, and the spiritual value of crusade and pilgrimage. This interest in worldly wisdom was continued by the Bavarian Neidhart von Reuenthal and again reflected by the popularity of Freidank's *Bescheidenheit* (*c.* 1230; "Modesty" or "Moderation") and Hugo von Trimberg's *Renner* (*c.* 1300; "The Runner").

Prose and drama. Medieval German prose literature was less substantial than that in verse. Berthold von Regensburg's sermons were eloquent vernacular works, and the writing of Mechthild von Magdeburg served as an early example of the mysticism that became important in the 14th century. The *Ackermann aus Böhmen* (*c.* 1400; *Death and the Ploughman*) by Johannes von Tepl marked the beginning of the humanistic tradition in Germany. The quality of Tepl's prose was unrivaled until the World War. Notable nonreligious works in Middle Low German were the *Sächsische Weltchronik* (*c.* 1235; "Saxon World Chronicle"), a history of the world compiled from Latin sources, and *Sachsenspiegel* (*c.* 1225; "Mirror of Saxony"), a compendium of medieval law written by Eike von Repgau.

Early works in drama—the fragmentary Easter play of Muri and the St. Gall Passion play (*c.* 1330)—led to a growing volume of popular morality and miracle plays, which were part of the church's campaign to instruct and inspire the lower orders. By the beginning of the 15th century, popular drama for secular entertainment appeared in the form of *Fastnacht,* or carnival, plays. These farcical comedies became increasingly popular, finally gaining respectability in the 1500s with the works of Hans Sachs.

Morality and miracle plays

From Middle High German to Baroque

The fall of Constantinople to the Turks (1453) created a fear of the Turk in western Europe that was often reflected in literature of the 16th and 17th centuries, while territorial rulers in Germany grew more independent at the expense of the empire. The invention of printing with movable type around 1440 revolutionized printing and literature. The old art forms continued, however, beyond 1440—medieval allegory, folk songs, songs of love and nature, and ballads survived in profusion. Hans Folz introduced more liberal rules to revive the art of the Meistersinger (members of guilds for cultivating singing and poetry). In the south, *Fastnachtsspiele* ("Shrovetide plays") became a vehicle of satire and broad humour, while in the north they were more restrained. Collections of comic anecdotes (*Schwänke*) evolved, often grouped around a single hero; anecdotes connected with Till Eulenspiegel, a 14th-century peasant jester, acquired a European reputation. From about 1450 a new bourgeois realism evolved.

THE COURTLY REVIVAL AND THE HUMANIST MOVEMENT

The Burgundian "renaissance." At some princely courts the Burgundian "renaissance" fostered a revived interest in medieval chivalry. At Innsbruck, Duchess Eleonore of Austria wrote a prose version of a French chivalrous romance, *Pontus und Sidonia* (1456); at Rottenburg, Jakob Püterich von Reichertshausen's work fostered a cult of Wolfram von Eschenbach. Medieval poems were printed, either in their original verse form or in prose versions of little literary merit. The allegorical works of the Holy Roman emperor Maximilian I, *Weisskunig* ("The White King") and *Theuerdank* ("Noble-Mind"), represented probably the last attempt to revive medieval chivalrous ideals.

Humanism. In sharp contrast to this courtly renaissance stood the new element in German literature after 1450: the humanist movement. The humanists between 1450 and 1480 drew their inspiration from Italy. Albrecht von Eyb, Heinrich Steinhöwel, Niklas von Wyle, "Arigo" (probably the pseudonym of Heinrich Schlüsselfelder), and Antonius von Pforr were chiefly translators who produced versions of many Latin classics, some Greek and Indian ones, and Italian works, including Boccaccio's *Decameron.*

German humanism took a different turn after 1480. Conradus Celtis, Eobanus Hessus, and others were closely associated with university circles and wrote almost entirely in Latin. A strong patriotic and political strain was evinced in the *Germania* (1501) of Jakob Wimpfeling. A lively interest in German history appeared in the work of Beatus Rhenanus, in the *Chronik der Äbbebte von St. Gallen* (1533; "Chronicle of the Abbots of St. Gall") of Joachim Vadianus (Joachim von Watt), and in the *Bayerische Chronik* (1533; "Bavarian Chronicle") of Aventinus.

The humanist impact was more direct in the drama.

Revitalization of drama by humanism

A lively dramatic tradition was established as part of university courses. Roman comedies and Latin plays by modern authors were performed for moral instruction and to teach eloquence. These plays contrasted sharply with the medieval dramatic types, and from them 16th-century German drama developed. Later this Latin drama was turned to the service of the Reformation.

The concentration on Latin by the most gifted writers of the age goes far to explain why vernacular literature between 1490 and 1520 was so scarce. Satiric or didactic works predominated; the most famous work of the 15th century, *Das Narrenschiff* (1494; "The Ship of Fools") by Sebastian Brant, reviewed all the vices of the age. In the Low German *Reinke de Vos* (1498; "Reynard the Fox"), verse was accompanied by a prose commentary applying the satirical episodes to contemporary evils.

Austrian literature under the Habsburgs. In Austria the vast expansion of the Habsburg empire brought the imperial chancellery from the Prague of the Luxembourg dynasty to the Vienna of the Habsburgs, carrying with it a double heritage: the German literary language and a humanist disposition and care for style. The emperor Maximilian I, himself a writer, made Vienna, through its chancellery and university, the leading humanist city in Germany with the aid of the scholars Conradus Celtis, Johannes Spiessheimer (Cuspinian), and Joachim Vadianus (von Watt).

Vienna acquired a theatre in the early years of the 16th century with the court plays of Celtis and the monastery plays of the abbot of the Schottenkloster ("Scottish," *i.e.*, Irish, monastery), Benedictus Chelidonicus. The German plays of the schoolmaster Wolfgang Schmeltzl were first performed in the Schottenkloster. From 1554 onward the Jesuits created out of the dynastic alliance of Austria and Spain a community of religious conviction and mental inclination. They began to perform a considerable repertoire of plays, and at the same time the Italian opera developed and was patronized especially by the emperors. The Viennese Baroque theatre became one of the artistic triumphs of Europe.

The individual regions flourished too. Tirol had its own courts from time to time. From 1430 there was an abundance of sacred and secular drama. A century later, in Sterzing, Vigil Raber was collecting and rewriting the plays of the people. About 1570 Archduke Ferdinand of Tirol led a poets' court, whose preacher was Johannes Nas, while the monk Laurentius von Schnüffis, from being theatrical producer to a duke, became a mystical lyricist. Salzburg had a life of its own, shining through a humanist, Paul von Hofhaimer, who set its tone, through the theatre of the archbishop's court and through the Benedictine university founded in 1620. Styria had a wide range of artistic interest in the great monasteries such as Kremsmünster, where Simon Rettenbacher wrote religious plays. Christoph van Schallenberg, Wolfgang Helmhard von Hohberg, and Katharina von Greiffenberg, poets who belonged to the nobility, were Protestants. It was the emperor Ferdinand II, sometime ruler of Styria, who decided the battle against the Reformation, first in Styria and then in his other lands. The voice of the triumphing church was uplifted in Vienna by Abraham a Sancta Clara, Habsburg court preacher and master of Baroque German prose.

THE REFORMATION

The Reformation and the vigorous protest of Martin Luther against ecclesiastical abuses affected life in Germany; it left little room for purely aesthetic considerations. The Humanists, who favoured reform without doctrinal changes, mostly held aloof, though with certain notable exceptions: Ulrich von Hutten, hoping for a political reformation too, supported Luther's cause in a series of pamphlets; Philipp Melanchthon, a Greek scholar, introduced the Humanist tradition into Protestant schools. But apart from the Humanists, almost every writer of note was preoccupied with the Reformation.

Literature of the Reformation

From the torrent of printed works that swept over Germany a few had literary merit: Luther's various writings, his eloquent Reformation pamphlets of 1520, for example, or his terrible condemnation of the peasants' revolt; *Die 15 Bundsgenossen* (1521–23; "Fifteen Comrades"), in which Johann Eberlin von Günzburg attacked religious abuses; Thomas Murner's satires in defense of the old religion; the fierce anti-Catholic satire in the *Fastnachtsspiele* of Niklaus Manuel of Bern; and the writings of Sebastian Franck, an independent radical thinker.

Luther's influence. In three respects the Reformation had a lasting effect on German literature. First, Luther's Bible translation (New Testament, 1522; Old Testament, 1534) was not only based for the first time on Hebrew and Greek texts but also achieved a vigorous, popular German style. It exercised an incalculable influence on the style and ideas of later German writers. Second, Luther established congregational hymn singing as an essential part of the Protestant service and wrote several hymns. His example established a tradition of hymn writing that has been a major contribution to the Christian world. Third, Luther had commended certain biblical subjects as suitable for plays, encouraging many dramatists—of whom Sixtus Birck, Paul Rebhuhn, Joachim Greff, Burkhard Waldis, and Jörg Wickram were the most outstanding—to write plays in German as vehicles of Lutheran teaching. Modeled on the Latin school drama, these plays marked a break with medieval types of drama.

Drama as a vehicle for Lutheran teaching

Swiss-German literature of the Reformation. The activities of the religious reformer Huldrych Zwingli had an indirect influence on literature. Zwingli himself wrote mainly in Latin. The so-called Zürich Bible of 1529 was gradually adapted to conform with the Luther Bible, which tightened the connection between Swiss and German writings. As a result of Zwingli's work, the Protestant majority of German-speaking Switzerland established a permanent connection with Protestant parts of western Switzerland and with Protestant countries abroad. The anonymous play about William Tell from the canton of Uri was a forceful and popular expression of Swiss patriotism. Gilg Tschudi's *Chronicon Helvetikum* (1734–36 "Swiss Chronicle"), covering the years 1000–1470 in Swiss history, endured as literature.

Other writings of the time. The Reformation did not account for all the vernacular literature after 1520. In scholarly works, such as the treatises of the Swiss doctor, alchemist, and scientist Paracelsus (Theophrastus von Hohenheim), German was beginning to replace Latin. But satirical or didactic works remained predominant, as, for example, the proverbs by Johann Agricola. The fable enjoyed a revival: Luther's translations from Aesop had no special merit, but Erasmus Alberus in his *Fabeln* (1534) and Burkhard Waldis in his *Esopus* (1548) turned the fable into a lively minor genre. *Grobianus* (1549) by Friedrich Dedekind, a satirical Latin guide to table manners, was translated into German by Kaspar Scheidt in 1551.

Apart from Luther, the most prolific and characteristic writer of the century was Hans Sachs of Nürnberg. His works, didactic in aim but entirely unpolemical, reflected ideals of the devout, upright, industrious Lutheran townsman and artisan. The greatest of the Meistersinger, he brought the form to perfection and virtually to an end, despite attempts at revivals by later poets. Sachs was at his best in comic verse anecdote and *Fastnachtsspiel*.

The works of Hans Sachs

As the tide of Reformation polemics receded, works of pure entertainment came into greater prominence. The popular chapbooks were often prose versions of medieval verse romances or were adapted from foreign sources. The stories of Fortunatus, Magelone, and Melusine were favourite reading far beyond the 16th century. The public also found distraction in collections of anecdotes. The *Schimpf und Ernst* (1522; "Jest and Earnestness") of Johannes Pauli had didactic aims, but entertainment was the sole aim of Jörg Wickram's *Rollwagenbüchlein* (1555; "Coach Book") and of similar collections by others. The modern novel began in Wickram's more substantial prose narrative—for example, *Der Goldfaden* (1557; "The Gold Thread")—and in the German translation (1569–95) of the Spanish chivalrous romance *Amadis de Gaula*.

THE CLOSE OF THE 16TH CENTURY

The Council of Trent (1545–63), which attempted to reform the Roman church, and the advent of the Catholic

Society of Jesus gave religious controversy a new turn. The Jesuits adapted the Latin school drama for educational purposes. Meanwhile, the Protestant academy in Strassburg, especially under the humanist Johannes Sturm, developed a rich educational and dramatic tradition.

In the last quarter of the century the most outstanding author was Johann Fischart, many of whose works were directed against the Jesuits, the Counter-Reformation, and notably a brilliant preacher, Johannes Nas. In all Fischart's writings, the didactic aims of the age were combined with an interest in literary form. *Das glückhafft Schiff* (1576; "The Ship of Good Fortune") was, formally and stylistically, one of the most distinguished poems of the century. Fischart's contemporary Philipp Nikodemus Frischlin was the last notable Latin dramatist of the century. On the other hand, the versions of Psalms by Paul Schede Melissus and Ambrosius Lobwasser, on the Calvinist model of Clément Marot, added a new element to the rich tradition of Protestant hymn writing, while an anonymous chapbook, *Historia von Dr. Johann Fausten* (1587), gave rise to the whole European Faust tradition. A noteworthy feature in the last years of the century was the arrival in Germany of troupes of English actors whose repertoire included versions of contemporary English plays. They influenced Jakob Ayrer and Heinrich Julius, duke of Brunswick.

Source of the Faust story

The 17th century

THE CHARACTER OF THE PERIOD

German 17th-century literature, widely known as Baroque literature, was a product of the times and of the social situation in Germany. Ushered in against a background of religious fervour and strife and political uncertainty, the early period was dominated by the Thirty Years' War and the shadow it cast over life. The local context was one of princely absolutism, each ruler emulating Louis XIV's court of Versailles and supporters of the Reformation struggling against those of the Counter-Reformation. New scientific discoveries were set in a society overshadowed by the influence of Luther, Machiavelli, Petrarch, and the Renaissance figures of the preceding century.

Baroque literature was dominated by writers who were guided by a sense of occasion rather than by the value of experience. They did not express themselves subjectively but used a medium of conventional images and strict formal patterns that heightened the emotional tensions of their themes. The themes were those common to European literature of the period and appeared in the lyric, epic, and drama: "fickle fortune" presides over the world, worldly pleasures are illusory, and man's duty is to fulfill his actor's role on the stage of life. Baroque literature presented these roles as examples of how man may escape the toils of fortune and cheat time: the martyr scorns the world and aspires to eternity, the statesman uses the present to his advantage, the stoic endures steadfastly, the hermit flees, while the pastoral figures escape to make-believe and the mystic seeks union with God.

MAJOR WORKS OF THE PERIOD

Lyric poetry. Lyric poetry quickly reached a high standard. In the *Buch von der deutschen Poeterey* (1624; "Book on German Poetry"), Martin Opitz demonstrated an elevated literary style in German, formulated rules such as the observance of stress in lines, and made other introductions and suggestions that he further demonstrated in his own poetry. His ideas were widely adopted and had a formative influence on later poets, including Georg Weckherlin and Paul Fleming.

Martin Opitz' influence

Given the basic similarity of many of its themes, the variety of Baroque verse is surprising. Paul Fleming's love poems and sonnets revealed emotional depth and lyric power, Andreas Gryphius was simultaneously pessimistic and religious, while formal dexterity in the patriotic poet and novelist Philipp von Zesen and verbal exuberance in the nature poetry of Johann Klaj illustrated an urge to experiment. Baroque poets wear masks: Gryphius regally expressed the existential problems of the day; Christian von Hofmannswaldau was a virtuoso of form who was intellectually flamboyant or soberly sincere according to his theme. In an age concerned with ultimate values, poets wrote both secular and religious verse, and in both this variety was present. The outstanding Lutheran poet Paul Gerhardt wrote hymns of quiet simplicity and warmth, the Jesuit Friedrich von Spee wrote religious pastoral lyrics, while Quirinus Kuhlmann was a mystic visionary.

Characteristics of Baroque poetry

The intellectual temper of the age found admirable expression in the epigram, which was developed to a fine art by Angelus Silesius and Friedrich von Logau. The variety of Baroque lyric poetry was matched by its volume.

Drama. In the first half of the century, vernacular drama was relatively uninspired. Later, with Daniel Caspar von Lohenstein, a dichotomy between excess of feeling and acuteness of intellectual perception was expressed in full-blown rhetoric. The most notable works of the period were those of Jakob Bidermann and Gryphius. Bidermann's Jesuit dramas dealt with themes of mutability, worldly vanity, and the urgency of salvation with great effectiveness. Gryphius' tragedies and comedies focused on stoicism and personal integrity. His work was literary rather than popular and so had little permanent influence, despite its artistic merit. The increasingly popular masques, ballets, and operas were to lay the foundations of the later, very different drama of Germany.

Prose narrative. The development of the novel was considerably influenced by Spanish, French, and neo-Latin sources. A host of foreign translations followed that of Mateo Alemán's *Guzmán de Alfarache* in 1615. Many were large, diffuse works, amalgams of moral and didactic elements with fantasy and amusement designed to satisfy the thirst for knowledge of the times. Some read like formal exercises in a particular form, but two writers stood out: the Austrian storyteller Johann Beer, who foreshadowed 18th-century realism, and Hans Jacob Christoph von Grimmelshausen, whose *Abenteuerlicher Simplicissimus* (1669; *The Adventurous Simplicissimus*) was one of the great novels of German literature, incorporating the principal themes of the age with metaphysical depth and religious insight.

The achievement of *Simplicissimus*

Philosophy and criticism. The Silesian mystic Jakob Böhme was outstanding as an original and influential philosopher whose doctrines inspired several religious separatist movements. Gottfried Leibniz' achievements in philosophy (see below) summed up the age and pointed toward the literature of the 18th century, when Baroque was to be rejected until its rediscovery in the 20th century.

The 18th century

THE AGE OF ENLIGHTENMENT

Rationalism. If religion was the dominant factor in German intellectual and spiritual affairs in the 17th century, the Enlightenment of the 18th brought about a reaction. Man now claimed to be able to understand the universe by virtue of his possession of the divine gift of reason. Empirical and idealist thinkers alike were united in rejecting traditional authority. In a rational universe, governed by the law of cause and effect, there was room neither for mystery nor for the doctrines of original sin and predestination. Evil was the result of irrational conditions of life, and man had it in his power to improve his lot by the pursuit of science and education. An optimistic belief in human perfectibility was generally held; it lay in the cultivation of reason and tireless effort in the service of human improvement. The man of the world being more highly regarded than the devout Christian, good taste and common sense came to be demanded, and literature assumed a markedly didactic character.

The foundations of rationalism had been laid by Leibniz. With him, the relationship of God and man to each other ceased to be considered within the limits of Christian dogma. German religious life was marked by a revival of Pietism, which left its traces in the sphere of religious poetry. The main emphasis lay not on conformity but on the individual's spiritual experience.

In literature the new ideas soon began to emerge. One of the most marked features of German literature in the 18th century was the progressive influence of English literature: first, of Joseph Addison, Jonathan Swift, Daniel Defoe,

The influence of English translations

and Alexander Pope; later of James Thomson, John Milton, and Edward Young. Translations and imitations of the English *Spectator, Tatler,* and *Guardian* helped to regenerate literary taste. Samuel Richardson had much effect upon the growth of the moral novel, while Young's *Conjectures on Original Composition* heralded a new epoch in German literature that was to be profoundly affected by the Scottish poet James Macpherson's *Ossian,* Bishop Thomas Percy's *Reliques of Ancient English Poetry,* and Shakespeare—the epoch of the Sturm und Drang ("Storm and Stress") movement.

The reaction against rationalism. Between 1724 and 1740 the critic Johann Christoph Gottsched succeeded in establishing in Leipzig literary reforms in accord with French 17th-century Classicism. He purified the stage and laid down principles according to which good literature was to be produced and judged. The limitations of Gottsched soon drew resistance from two Swiss scholars, Johann Jakob Bodmer and Johann Jakob Breitinger. Basing their arguments on John Milton's poem *Paradise Lost,* they insisted that imagination should not be dominated by reason. The effects of the controversy appeared toward midcentury in a group of Leipzig writers of Gottsched's own school, the Bremer Beiträger (Bremen Contributors), as they are usually called after the paper in which they published their work. In this—the *Neue Beiträge zum Vergnügen des Verstandes und Witzes*—there appeared in 1748 the first installment of an epic by Friedrich Gottlieb Klopstock, *Der Messias* (completed 1773), whose theme created a sensation when the first cantos appeared. Klopstock's genius was, however, more suited to the lyric, and his odes, in which sentimental and patriotic themes were prominent, were much admired. Friedrich von Hagedorn showed to what perfection occasional verse could be brought, while Ewald Christian von Kleist excelled in sentimental nature poetry. Meanwhile, a rising interest in Germanic antiquity aided the growth of the "bardic" movement led by Heinrich Wilhelm von Gerstenberg, Karl Friedrich Kretschmann, and Michael Denis, the translator of Macpherson's *Ossian.* A notable group of poets was the Göttinger Hain (Göttingen Grove) founded in 1772. Johann Heinrich Voss, the group's leader, was author of the famous idyll *Luise* (1795).

The influence of Lessing. As Klopstock had been the first of modern Germany's inspired poets, so Gotthold Ephraim Lessing was the first critic who brought credit to the German name throughout Europe. Like Gottsched, he had unwavering faith in Neoclassicism, but classic meant for him, as for his contemporary Johann Joachim Winckelmann, Greek art and literature rather than French pseudo-Classicism, though it is true that Lessing's own exposition of Aristotle's theory of tragedy was full of the moral preoccupations of the Enlightenment. He looked to England rather than to France for the regeneration of the German theatre. His own dramas were pioneer works in this direction. *Miss Sara Sampson* (first performed 1755) was a bourgeois tragedy on the English model; *Minna von Barnhelm* (1767), a comedy in the spirit of George Farquhar; in *Emilia Galotti* (1772) Lessing remodeled the "tragedy of common life" in a form that came to be acceptable to the Sturm und Drang; and finally in *Nathan der Weise* (1779; *Nathan the Wise*) he won acceptance for iambic blank verse as a medium for elevated drama.

Lessing's dramatic works

Because of Lessing, German literature made a great leap forward beyond the feeble achievements of the first half of the century. The domestic tragedy, its plot centred on the problem of class distinction, foreshadowed plays involving marked political and social criticism in the Sturm und Drang period, while *Nathan* was a forerunner of the "drama of ideas" of Weimar Neoclassicism. Lessing's theoretical work placed criticism in the forefront of affairs in literary Germany. His sharp rejection of descriptive poetry had a great effect on the writing of the next generation, and his attack on French literary authority prepared the way both for a greater attention to English examples and for the search for native originality.

Wieland's contributions

Christoph Martin Wieland contributed to the widening of the German imagination by introducing remote and exotic settings. With the exception of his verse-romance *Oberon* (1780), his work fell into neglect; he did excellent service, however, to the development of German prose fiction with his psychological novel *Agathon* (1766–67; *The History of Agathon*) and his satire *Die Abderiten* (1774; "The Abderites"; Eng. trans., *The Republic of Fools . . .*). The German novel owed much to the example of *Agathon,* but its groundwork and form were borrowed from English models.

THE AGE OF GOETHE

Sturm und Drang. The period of Neoclassicism and Romanticism, the greatest epoch in German literature, fell within the lifetime of Johann Wolfgang von Goethe. The age of Goethe went beyond the Enlightenment's substitution of science for religion, inasmuch as it ascribed to science only a peripheral position in relation to the ultimate questions of life. It insisted upon the value of feeling in face of the limitations of reason. Impulse, instinct, emotion, fancy, and intuition acquired a quasi-religious significance as being the links that connect man with divine nature. The ideal of the classical age, soon to be called *Humanität* ("humanness"), was that of the fully developed personality in which intellect and feeling should be harmoniously balanced. Three phases may be distinguished in the evolution of this new outlook: Sturm und Drang, Classicism, and Romanticism.

Goethe belonged to and profoundly affected the Sturm und Drang movement, which aimed at overthrowing rationalism. Seeds of the new growth were to be found in Klopstock, in the spiritual force of Pietism, and in the rising resistance to French Classical taste, while the influence of Rousseau, Young, Macpherson, and the recently translated Shakespeare was of prime importance. Nature, genius, and originality were the slogans of the new movement, and an increasingly oppressive sense of dissatisfaction with the civilization of the day assailed its exponents. The cult of nature replaced orthodox religion. No law was recognized as being above the individual conscience. The standard outlook thus demanded unceasing effort, like that of Faust. Strain, protest, revolt, yearning, disillusion were obvious on all sides, and egotism became a dominating feature in literature and thought.

The influence of Herder

The critical writings of Heinrich Wilhelm von Gerstenberg stressed personal feeling in matters of taste, but the chief impetus came from Johann Georg Hamann, who emphasized the inspirational and symbolical function of language. His pupil was Johann Gottfried von Herder, who grasped, as no thinker before him had done, the idea of historical evolution and engendered the main current of the Sturm und Drang. He stressed the value of historical continuity in literature and pointed to the folk songs, ballads, and romances of the Middle Ages as sources of inspiration to which Bishop Thomas Percy's *Reliques of Ancient English Poetry* (1765) had recently drawn attention.

It was, moreover, Herder who aroused in Goethe an interest in Gothic architecture, the *Volkslied,* and Shakespeare. A pamphlet, "Von deutscher Art und Kunst" (1773; "Concerning German Nature and Art"), was a kind of manifesto of the Sturm und Drang. The new ideas seemed at once to set Goethe's genius free. His *Götz von Berlichingen* (1773), the first important drama of the Sturm und Drang, was followed by the first novel of the movement, *Die Leiden des jungen Werthers* (1774; *The Sorrows of Young Werther*), which made the author famous throughout Europe. In all forms of literature he set the fashion for his time. The Shakespearean restlessness of *Götz* found imitators in Jakob Michael Reinhold Lenz, Friedrich Maximilian von Klinger, Johann Anton Leisewitz, Heinrich Leopold Wagner, and Friedrich Müller. The dramatic literature of the Sturm und Drang was its most characteristic product; it was inspired by a desire to present upon the stage figures of Shakespearean grandeur impelled by gigantic passions, all considerations of plot, construction, and form being subordinated to character, and all accepted authority—literary, social, political, or moral—being rejected.

With the production of *Die Räuber* (1781; *The Robbers*) by Friedrich Schiller, the drama of the Sturm und Drang entered a new phase. Schiller's tragedy was more skillfully

The dramas of young Schiller

adapted to the exigencies of the theatre than those of his predecessors had been; it and the succeeding dramas, *Die Verschwörung des Fiesko zu Genua* (1783; *Fiesco; or the Genoese Conspiracy*) and *Kabale und Liebe* (1784; *Cabal and Love*), were fine pieces of high promise. Germany owed its national theatre to the Sturm und Drang period; permanent theatres were established in these years at Hamburg, Mannheim, and Gotha, and the Burgtheater (now the Hofburgtheater) was founded at Vienna in 1776 by the emperor Joseph II.

Neoclassicism. Herder's doctrine of *Humanität,* in which intellect and feeling are reconciled, became fundamental to German Neoclassicism. This harmonious balance and self-discipline was lacking in the Sturm und Drang, and the movement soon exhausted itself. A more positive form of moral idealism appeared in the poetry of Goethe and Schiller, as well as in the philosophy of Immanuel Kant. The problem of freedom was rendered acute by the impact of the French Revolution, which German literature generally regarded as a warning. Schiller believed the antagonism between duty and inclination could be resolved once morality became "second nature," and this could be achieved only through the contemplation and production of beauty. Art thus acquired an educational function, and aesthetic education was one of the major objectives of Neoclassicism.

Moral idealism

Goethe's Weimar period. For Goethe a new phase in his development began with his departure for Weimar in 1775, while after *Don Carlos* (1787) Schiller turned from poetry to study history and philosophy. The first 10 years of Goethe's life in Weimar were marked by his renewed friendship with Herder, by his public service as a minister of state, and by his emotional attachment to Charlotte von Stein. He did not achieve greater clarity in his ideas until after his sojourn in Italy (1786–88). In Italy he turned his attention to three dramatic works; he gave *Iphigenie auf Tauris* (*Iphigenia in Tauris*) its final form (1787), completed *Egmont* (1788), and replanned *Torquato Tasso* (1790). *Wilhelm Meisters Lehrjahre* (*Wilhelm Meister's Apprenticeship*), Goethe's most important novel, had become, by the time it appeared in 1795–96, a book on the conduct of life. It is an outstanding example of the bildungsroman ("educational" novel), a characteristic German novel form, and profoundly affected future practitioners of the genre.

Before *Wilhelm Meister* appeared, however, German philosophical thought and literature had arrived at that degree of stability in form and ideas essential to a great literary period. In the year of Lessing's death (1781), Kant had published his *Kritik der reinen Vernunft* (*Critique of Pure Reason*). Under the influence of Kant, Schiller turned to the study of aesthetics, the first fruits of which were his philosophical lyrics and his treatises "Über Anmut und Würde" (1793; "On Grace and Dignity") and, between 1795 and 1796, *Briefe über die ästhetische Erziehung des Menschen* (*Letters on the Aesthetic Education of Man*) and "Über naive und sentimentalische Dichtung" ("On Naive and Sentimental Poetry").

Goethe and Schiller. The years 1794 to 1805, when Goethe and Schiller were united in close friendship in Jena and Weimar, mark the culmination of literary Neoclassicism. Schiller provided the theoretical basis; Goethe, as director of the ducal theatre, influenced the whole practice of dramatic production in Germany. Under his encouragement Schiller turned from philosophy to poetry and between 1798 and his death in 1805 wrote a series of Classical dramas that are Germany's greatest: the *Wallenstein* trilogy, *Maria Stuart, Die Jungfrau von Orleans* (*The Maid of Orleans*), *Die Braut von Messina* (*The Bride of Messina*), and *Wilhelm Tell,* closing with the fragment *Demetrius.*

The universality of the greatest writers is also reflected in others: Karl Philipp Moritz wrote on aesthetics and mythology and produced the novel *Anton Reiser* (1785–90); the physicist Georg Christoph Lichtenberg wrote on the English artist William Hogarth and was Germany's greatest master of the aphorism.

Goethe's *Faust*

The supreme work of Goethe's later years is *Faust,* Germany's greatest contribution to world literature. Part I (1808) sets out Faust's despair, his pact with Mephistopheles, and his love for Gretchen; Part II (1832) covers the magician's life at court, the winning of Helen of Troy, and Faust's purification and salvation. The doctrine of the fulfillment of life by striving and selfless activity revealed in *Faust* was fundamental to Goethe's mature works. The tragic novel *Die Wahlverwandtschaften* (1809; *Kindred by Choice*) insisted on the theme of renunciation. *Wilhelm Meisters Wanderjahre* (1821–29; *Wilhelm Meister's Travels*), with its social utopianism and teaching of restraint, offered a criticism of the rise of industrialism. The autobiographical *Dichtung und Wahrheit* (1811–33; *Poetry and Truth: From My Own Life*), containing dramatic pieces, scientific writings, and lyrics, indicates the many-sidedness of Goethe's achievement (see GOETHE).

THE ROMANTIC MOVEMENT

First phase. The Romantic Movement began less as a protest against the Neoclassicism of Weimar than as a radical extension of some of its beliefs and interests, especially, at first, in its emphasis upon Greek antiquity, longed for like some lost paradise. The Romantic poet could create his own world from reality or from fancy and could turn whatever he liked into poetry. There was to be no end to the innovations made in content and style by the great wealth of literary talents who now emerged all over Germany and from various strata of society. The rising generation felt free and able to revise all accepted representative values, not only in art and literature but in other spheres as well. Among the topics then in vogue were nature and the spirit in all their manifestations, particularly the supernatural, the subconscious, and the mystical. In the evolution of German Romanticism no small part was played by the philosopher Johann Gottlieb Fichte and the theologian Friedrich Schleiermacher.

Two major writers fall between Neoclassicism and Romanticism proper: Friedrich Hölderlin and Jean Paul (Johann Paul Friedrich Richter). Hölderlin was one of Germany's greatest poets. He was a friend of the philosophers Friedrich Schelling and G.W.F. Hegel and was influenced by Klopstock and Schiller. His lyrical novel *Hyperion* (1797–99) sums up his major concerns: his yearning for antiquity, for union with the divine, and for a political renewal of Germany. His poems in Neoclassical metres and free rhythms convey a vision of sublime nobility. Jean Paul's once immensely popular novels introduced a new focus on ordinary life. They lack shape but sustain interest by their display of humour, warmth, sentiment, and whimsy. His main novels include *Hesperus* (1795), *Siebenkäs* (1796–97; *Flower, Fruit and Thorn Pieces*), *Titan* (1800–03), and *Die Flegeljahre* (1804–05; *Walt and Vult*).

Hölderlin and Jean Paul

The first Romantic school originated in Jena about 1798. It was partly inspired by the subjective idealism of Fichte, but its principal philosopher was Schelling, whose *Naturphilosophie* asserted the unity of nature and the human spirit. The major literary theorists were the brothers August Wilhelm and Friedrich von Schlegel, who held that the first duty of criticism was to understand and appreciate, while Romantic literature was to encompass all forms of writing in "progressive universal poetry." Their main literary model was Goethe's *Wilhelm Meister.* The chief creative writers of the Jena school were Wilhelm Heinrich Wackenroder, Ludwig Tieck, and Novalis (Friedrich von Hardenberg). Wackenroder's collection of anecdotal accounts and sketches, *Herzensergiessungen eines kunstliebenden Klosterbruders* (1797; "Effusions of an Art-Loving Friar"), was the school's first major literary production, and it gave to art a religious significance. Tieck was a skillful and prolific prose writer and a tireless publicist. The finest imaginative achievement of early Romanticism, however, was found in Novalis' lyrics and aphorisms and in his unfinished novels, notably *Heinrich von Ofterdingen* (1802; *Henry of Ofterdingen*). These works combined abstract ideas with symbols of beauty and innocence. Romanticism's universal sympathies were also exemplified by many superb translations, of which the greatest were the translations of Shakespeare's plays by August Wilhelm von Schlegel and Tieck.

Wackenroder, Tieck, and Novalis

The second Romantic school. The first Romantic school had dispersed by 1804. After 1805, however, a second school developed in Heidelberg around Achim von Arnim, Clemens Brentano, and Johann Joseph von Görres. Unlike the members of the earlier school, the Heidelberg writers produced historical works and also collected folk songs and popular prose romances. From this period dates the scholarly study of German philology and medieval literature. The most characteristic production was the folk-song collection published by Arnim and Brentano under the title *Des Knaben Wunderhorn* ("The Youth's Cornucopia") in 1805–08. The same impulse later led the brothers Wilhelm and Jacob Grimm to compile their famous collection of fairy tales, *Kinder- und Hausmärchen* (1812–15; "Children's and Household Stories"; Eng. trans. *Fairy Tales*).

Interest in folk heritage

The folk song had a profound influence on Romantic poetry and not least on its greatest exponents, Brentano and Joseph von Eichendorff. Brentano's poetry, collected posthumously in 1852–55, is characterized by intense musicality. Eichendorff's lyrics, collected in 1837, nostalgically describe man, God, and nature in simple but highly evocative imagery.

The concept of universal poetry encouraged a fashion for short poetic fiction that combined the realism of the *Novelle* form with fairy-tale fantasy, often interspersed with lyrics. Illustrative examples include Tieck's *Der blonde Eckbert* (1797), Friedrich de la Motte Fouqué's *Undine* (1811), Adelbert von Chamisso's *Peter Schlemihls wundersame Geschichte* (1814; *The Wonderful History of Peter Schlemihl*), and Eichendorff's *Aus dem Leben eines Taugenichts* (1826; *Memoirs of a Good-for-Nothing*).

The Heidelberg school broke up about 1809, and Berlin and Dresden became the main centres of Romanticism. Two major writers are associated with this phase, Heinrich von Kleist and E.T.A. Hoffmann. Both explored the darker aspects of life that had always interested the Romantics.

Kleist and Hoffmann

Kleist was a master of the *Novelle* form (*Erzählungen* [1810–11; "Tales"; Eng. trans. *The Marquise of O. and Other Stories*]) and the greatest dramatist of his day. His plays include *Amphitryon* (1807), *Penthesilea* (1808), and *Prinz Friedrich von Homburg* (posthumously published, 1821). In such works, Kleist depicted the tragic impossibility of understanding a world riddled with ambiguities. Hoffmann, a storyteller of genius, gave European currency to German Romanticism in countless works of fantasy and the grotesque. Among his tales are *Der goldne Topf* (1814; *The Golden Pot*) and *Das Fräulein von Scudéri* (1819; *Madame de Scudéri*). Other figures of note in this phase were the dramatist Zacharias Werner and the poet Wilhelm Müller, whose song cycles such as *Die Winterreise* (1824; "The Winter Journey") were set to music by Franz Schubert.

Romanticism entered its final phase with the Swabian school, whose more talented members included Ludwig Uhland, Justinus Kerner, and Gustav Schwab. By this time Romanticism had become stereotyped and lacked its original vigour.

The 19th century

The death of Goethe in 1832 marked the end of the cosmopolitan humanism typical of the 18th century. Although some leading Romantics outlived Goethe by a decade, the movement had lost its impact. In conservative, nonliterary fields, Romanticism was more tenacious, and in politics its alliance with rising nationalism coloured German thinking for more than a century. In literature, Goethe himself was of more lasting significance; few writers were unaffected by his work, though the effect sometimes took the form of rebellion against his Olympian predominance.

Rule by conservative governments repressed liberty of thought, and writers' efforts to prescribe solutions for social ills were foiled by censorship. Literature was dominated by disillusionment with man's capacity to achieve lofty ends, and pessimistic appraisal of man's role replaced once optimistic or constructive attitudes. In keeping with this change in attitude, writers sought to free themselves from the bondage of Neoclassical and Romantic thought, not always by rejection but often by adaptation.

GRILLPARZER, BÜCHNER, AND THE DRAMA

The post-Napoleonic era produced several fine dramatists, notably the Austrian Franz Grillparzer and the German Georg Büchner. Grillparzer consciously formed his dramas in the mold of Neoclassicism but filled them out with a new vein of realism and a theatricality that he derived from the Viennese popular theatre. His tragedies on Classical and historical subjects had as their theme the conflict between active life and contemplative life. *Sappho* (1818) was modeled on Goethe's *Tasso*, and *Das goldene Vliess* (1820; *The Golden Fleece*) was a tragedy of stark psychological realism. The historical plays mirrored contemporary events: *König Ottokars Glück und Ende* (1823; *King Ottocar, His Rise and Fall*) reflects Napoleon's fate, while *Ein Bruderzwist in Habsburg* (*c.* 1848; *Family Strife in Hapsburg*) conveys Grillparzer's pessimism about the possibility of right action in mid-19th-century politics.

Grillparzer's historical plays

Whereas Grillparzer was politically and poetically a conservative, Büchner was in every respect the opposite. He turned to literature after failing to start a revolution in his native Hessen. His plays were rooted in the achievements of the Sturm und Drang but anticipated 20th-century dramatic forms. In *Dantons Tod* (1835; *Danton's Death*), Büchner portrayed the failures of the French Revolution, and in *Woyzeck* (posthumously published in 1879) he created the first major tragedy with a lower-class hero. Büchner viewed man pessimistically but with compassion as a victim of social, historical, and other forces: mere existence causes pain, and the cosmos is a spiritual void.

Significance of Büchner's *Woyzeck*

Büchner and Grillparzer each wrote one comedy, but the most successful comic dramatists were Ferdinand Raimund and Johann Nepomuk Nestroy, who gave lasting currency to the Viennese popular stage. The ambitious tragedies of Christian Dietrich Grabbe, such as *Napoleon* (1831), have not stood the test of time, but he also produced one memorable comedy in *Scherz, Satire, Ironie und tiefere Bedeutung* (1827; *Satire, Irony, and Deeper Meaning*). The heritage of Neoclassicism and Romanticism had been discredited, but no new faith emerged from the resulting disillusionment.

LYRIC POETRY

The great tradition of Neoclassical and Romantic poetry made it possible to achieve formal excellence but difficult to be original. Friedrich Rückert and August von Platen-Hallermünde, two consummate formal lyricists, illustrated this amply. A greater depth of feeling, however, was achieved in the melancholy lyrics of Nikolaus Lenau.

For Eduard Mörike, too, classical poetry was a model; Mörike's formal excellence was matched by an idyllic and melodious portrayal of nature and country life. In *Mozart auf der Reise nach Prag* (1856; *Mozart on the Way to Prague*), he humorously examined the problems of artists in a world uncongenial to art. Nature was also a source of inspiration for Annette von Droste-Hülshoff, whose powerful rhythm and sombre language expressed apprehension of the irrational forces of life, as in *Das geistliche Jahr* (1851; "The Spiritual Year"). Religious feeling helped balance her vision and gave her poetry greater maturity.

Heine. No one attacked Romanticism more ruthlessly than Heinrich Heine. In Germany, Heine was, and still is, a controversial figure, mainly because he subjected German national susceptibilities and Romantic nationalism to scathing criticism, but his *Buch der Lieder* (1827; "Book of Songs") became one of the best known anthologies of love poetry. Heine described dreams and yearnings, and his realism showed that they were only make-believe. In *Romanzero* (1851) and the posthumously published poems, his poetry conveyed the hopes and anguish that were so real during his last long, drawn-out illness, and his early Saint-Simonian belief in the "rehabilitation of the senses" that had given way to a belief in God. His lesser-known prose, *Reisebilder* (1826–31; "Travel Sketches"); *Lutezia* (1854), a collection of reports on life in France; his analysis of German intellectual life and history in *Zur Geschichte der Philosophie und Religion in Deutschland*

Heine's attack on Romanticism

(1834; "On the History of Religion and Philosophy in Germany"), reveal him as a master of ironic prose, dissatisfied with solemnity and pretension. His most effective political satire is a verse epic, *Deutschland: Ein Wintermärchen* (1844; "Germany: A Winter's Tale"), a savage attack on personal enemies and political conditions of his time.

"Young Germany." In 1835 an edict of the federal Diet banned Heine's writings, together with those of Ludolf Wienbarg. Karl Gutzkow, Heinrich Laube, and Theodor Mundt. Wienbarg's *Ästhetische Feldzüge* (1834; "Aesthetic Campaigns") had been dedicated to "Young Germany," and this name was then given to the writers who were in tune with the radicalism of the Young Hegelians and political liberals, frustrated by severe censorship and authoritarian government. Wienbarg advocated a literature to deal with political and social problems. Karl Gutzkow criticized conventional morality and orthodoxy in a novel, *Wally, die Zweiflerin* (1835; "Wally the Doubter"), and a drama, *Uriel Acosta* (performed 1846). Exile was often the fate of those who dared to criticize the established political and social order. Ludwig Börne, a highly talented prose writer, went into self-imposed exile in Paris, and his *Briefe aus Paris* (1830–33; "Letters from Paris") were valuable social documents. Two important lyric poets, Georg Herwegh and Ferdinand Freiligrath, had to flee, one to Switzerland, the other to London.

Literature of socio-political commitment

More conservative was Emanuel Geibel, whose collections, such as *Zeitstimmen* (1841; "Voices of the Age"), contain patriotic verse. He was the leader of the popular Munich school of poetry, whose patron was Maximilian II of Bavaria. Of these poets only Heinrich Leuthold struck a deeper note. Paul von Heyse's novellas rarely roused the reader, and Victor von Scheffel's verse tale *Der Trompeter von Säckingen* (1854) and novel *Ekkehard* (1857), though popular at the time, came to sound unconvincing.

REALISM AND REGIONALISM

Characteristics of Poetic Realism

Realism, often of regional inspiration, was a source of originality at this time. Its finest exponents were Adalbert Stifter, Gottfried Keller, and Theodor Fontane. Poetic Realism, a term coined by Otto Ludwig, aimed at portraying life but only insofar as life was artistically significant and appeared to possess intrinsic value. Attention focused on social reality but not, as in later naturalism, on its ugly, pathological side. The main objective of a realist writer was to discover positive values in everyday life without reference to transcendental ideas. Changes in the social order had caused a host of social critics to question developments that accompanied the beginning of urbanization. Of them Karl Marx is the best known, but Arthur Schopenhauer's pessimistic philosophy and Ludwig Feuerbach's materialistic thought also amounted to a more sober appraisal of man's capacity. There also developed a positivism that, by way of analogy, sought to apply to the study of literature and society methods that were mistakenly believed to be those of natural science. This, in turn, led to the study of sources and texts, formalized by the first important organized school of literary history in Germany, that of Wilhelm Scherer. Together with his successors, Erich Schmidt and Jakob Minor, Scherer established criticism of modern literature as an academic discipline; the Grimms and Karl Lachmann had already given academic respectability to German medieval studies.

In fiction, writers concentrated on subjects that they could, through familiarity, accurately describe. Karl Immermann, whose work was still greatly influenced by Neoclassicism, in *Der Oberhof* (1839; "The Manor") portrayed peasants rooted in their work and their countryside. The Low German novels *Ut de Franzosentid* (1859; "During the Time of the French Conquest"; Eng. trans. *When the French Were Here*) and *Ut mine Stromtid* (1862–64; "During My Apprenticeship"; Eng. trans. *Seed-time and Harvest*) by Fritz Reuter contained a wealth of individual character made more convincing by a lively dialect style. With *Quickborn* (1853), a collection of lyrical poetry, Klaus Groth became a prototype of the regional poet; his dialect clearly linked his work to colloquial speech and recalled the folk song. A major dialect writer was the Swiss novelist Jeremias Gotthelf (Albert Bitzius). A close knowledge of the life of the Swiss peasants and their problems is reflected in his novels, of which *Uli der Knecht* (1846; *Ulric the Farm Servant*) and *Uli der Pächter* (1849; "Ulric the Tenant-Farmer") are the best known. Concerned for the moral welfare of the peasants, he preaches against liberalism in politics and against the loosening of moral sanctions and seeks to advocate a life of probity based on communal responsibility. His works convince the reader because of his shrewd insight into the mind of the peasants, his realistic assessment of their motives, and his faithful description of the peasant community. In *Die schwarze Spinne* (1842; *The Black Spider*) the events and persons, though realistically described, assume almost symbolical importance.

Adalbert Stifter, too, drew strength from his native Bohemian forest; some of his tales, collected in *Studien* (1844–50) and *Bunte Steine* (1853; "Colourful Stones"), are set there, but his language is Classical, reflecting his quest for stylistic perfection. For Stifter the world of everyday events is a symbol of emotional significance; he carefully portrays it and can thus be called a Poetic Realist. In *Nachsommer* (1857; "Indian Summer"), a bildungsroman influenced by Goethe's *Wilhelm Meister,* Stifter stresses the power of art to educate; he seeks to show how the gentle law of humane action, based on justice, simplicity, self-control, restricted activity, and admiration of the beautiful, is effective in bringing about an exemplary life true to nature. Renunciation of violence is the major theme of *Witiko* (1865–67), a historical novel about the growth of culture in the 12th century; humane restraint is also the message of his story *Die Mappe meines Urgrossvaters* (1841–67; "The Portfolio of My Great Grandfather").

The gentle art of Stifter

The realism of Otto Ludwig had a psychological flavour. "Die Heiterethei" ("The Cheerful Ones") and "Aus dem Regen in die Traufe" ("From the Frying-Pan into the Fire"), the two tales making up his collection *Thüringer Naturen* (1857), are a humorous exploration of life in his native land, while in his novel *Zwischen Himmel und Erde* (1856; *Between Heaven and Earth*) a conflict within an artisan family is explored with striking objectivity and careful characterization. His play *Erbförster* (1849; *The Forest Warden*) is a domestic tragedy of a forester obsessed by a sense of justice who finally shoots his own daughter in mistake for the son of his enemy. While Ludwig's *Shakespeare-Studien* (1871) reveals a fine understanding of dramatic art, his own *Die Makkabäer* (1854; "The Maccabees") is a failure.

With Gottfried Keller a pinnacle of Poetic Realism in prose narrative is reached. The scene of his works is his native Switzerland. All his writings reveal his attempts to differentiate between those characters whose thought and conduct allow their personalities to mature and those who, ignoring the voice of nature, fail to develop their inner potentialities. *Der grüne Heinrich* (1854–74; *Green Henry*), a semiautobiographical bildungsroman, is the story of the lively struggle and development of a Swiss painter. Keller portrays a romantic personality seeking to come to terms with life and shows how youthful dreams may become mutilated and how the artist's vision has to be readjusted to the demands of everyday life. *Die Leute von Seldwyla* (1856–74; "The People of Seldwyla"), a collection of *Novellen,* reveals his humour at its best. He resists the flights of a romantic imagination and cautiously consolidates his appraisal of everyday life. *Martin Salander* (1886) is political in tone, and its strictures on the liberalism of the day do not enhance its artistic value.

The underlying theme of Keller's works

Another important realist was Theodor Storm, whose work is stamped by the atmosphere of his native Schleswig-Holstein. In his work romantic elements were gradually subordinated to realistic description. The elegiac, often sentimental, tone of his earlier writing prevailed less and less, though both his prose tales and his lyric poetry were permeated by his sense of the ephemerality of life. Romantic preoccupation with the past had stimulated historical thought and writing. Some of Storm's and Keller's *Novellen* had dealt with the past, but other writers, such as Willibald Alexis, Conrad Ferdinand Meyer, Wilhelm Hauff, Gustav Freytag, and Wilhelm Heinrich Riehl, made history their main theme. The stories of Alexis are

imbued by a delicate sense of humour and a feeling for the landscape of Brandenburg. Freytag was less successful in his historical novel *Die Ahnen* (1872–81; "The Ancestors") than in portrayal of social and economic changes of his age in *Soll und Haben* (1855; *Debit and Credit*) and in his comedy *Die Journalisten* (1854). Friedrich von Spielhagen described social conditions more amply in a series of novels written after 1861. Though such works contain political criticism, they are spoiled by their sentimentality.

Wilhelm Raabe's analysis of social life is more profound. His writings appear complex; he anticipates 20th-century methods of storytelling in focusing attention not only on the story but on the way in which the story is told. He attacked the narrowness of the bourgeois philistinism and the nationalism of Otto von Bismarck's German empire. His humour helped him to overcome the pessimism of his early work, of which *Der Hungerpastor* (1864; *The Hunger Pastor*), *Abu Telfan, oder Die Heimkehr vom Mondgebirge* (1868; *Abu Telfan, Return from the Mountains of the Moon*), and *Der Schüderump* (1870; "The Rickety Cart") are striking examples. In his later work—*e.g., Alte Nester* (1880; "Old Nests") and *Stopfkuchen* (1891; "Cake Eater")—he depicted eccentric characters with a rich inner life who achieve spiritual freedom. His pessimism and humour are paralleled by Wilhelm Busch, whose laughter over human imperfection savagely exposes hypocrisy and illusion.

In Austria gentler moods prevailed during the late 1800s. Ferdinand von Saar and Marie Ebner von Eschenbach provided realistic accounts of both bourgeois and peasant Austrian society. Peter Rosegger and Ludwig Anzengruber also wrote about peasant life.

Meyer's fusion of poetry and reality

Conrad Ferdinand Meyer, a Swiss, through his understanding of history, achieved a rare fusion of poetry and realism. He described in chiseled prose the downfall of great men through the lust for power. The best of his *Novellen*, such as *Jürg Jenatsch* (1874), were inspired by an insight into the life of the Renaissance that had been stimulated by the work of the Basel historian Jacob Burckhardt. His lyrical poetry, like his prose, showed a rare sense of form. It is symbolical poetry that subordinates his personal feeling to imagery.

The plays of Friedrich Hebbel revealed Poetic Realism at its most powerful. His work was a synthesis of psychological analysis and metaphysical beliefs. Deep inner impulses drove his characters to doom: his dramas, *Judith* (1841), *Herodes und Mariamne* (1849), and *Gyges und sein Ring* (1856; *Gyges and His Ring*), depicted the tragedy of those who suffered defeat because their outraged individuality did not allow them to compromise. In his last play, *Die Nibelungen* (performed 1861), he interpreted an old legend in terms of his own psychological and metaphysical ideas. Theodor Fontane produced the first true social novels in German. His realism was subtle and impressive for its humour and irony, but it also contained a strong poetic vein. In such novels as *L'adultera* (1882), *Irrungen Wirrungen* (1888; "Trials and Tribulations"), *Frau Jenny Treibel* (1892), and *Effi Briest* (1895) he combined psychological insight with an understanding of social conditions: human relationships clashed with society and survived or broke down according to their innate strength. Traditional social codes proved inadequate, but new ones had yet to be forged.

The social novels of Fontane

Nietzsche's theory of art

Friedrich Nietzsche was a harbinger of 20th-century literature. His distinction, in *Die Geburt der Tragödie aus dem Geiste der Musik* (1872; *The Birth of Tragedy*), between the Apollonian and Dionysian elements of art was of considerable consequence. The view spread that Classical art could not only be serene but also could be ecstatic, and that the origins of Greek drama sprang from the orgiastic intoxication of Dionysian religious mysteries. Nietzsche's emphasis on a need to liberate personality from the shackles of conventional Christian morality (he denounced Richard Wagner after the composer had turned to Christianity), his skepticism as to the validity of the artist's statements and his place in society, and his prophecy of the nihilism to come provided an arsenal of ideas and intellectual ferment for the next generation of writers (see NIETZSCHE).

NATURALISM

The keynote of naturalism in Germany was scientific objectivity, and the principal model was the work of Émile Zola in France. An anthology of lyric verse, *Dichtercharaktere*, appeared in 1884, in which urban life was the theme, but the real revolution was made by Arno Holz, who, in *Buch der Zeit* (1886; "Book of the Times"), revealed himself as the first important poet of naturalism. Together with Johannes Schlaf, he wrote three tales published as *Papa Hamlet* (1889), in which they attempted faithfully to depict the minutiae of life, even its pathological and sordid aspects. Gerhart Hauptmann was the chief naturalist playwright. His play *Vor Sonnenaufgang* (1889; *Before Dawn*) was memorable for its novel technique: it was a drama without hero or proper plot. *Die Weber* (1892; *The Weavers*) was an indictment of dire poverty caused by industrialization, while *Der Biberpelz* (1893; *The Beaver Coat*), one of the few successful German comedies, was a satire on Prussian officialdom. In *Hanneles Himmelfahrt* (1893; "The Assumption of Hannele"), Hauptmann began to experiment with Symbolist drama, and *Der Narr in Christo, Emanuel Quint* (1910; *The Fool in Christ*) revealed his force as a prose narrator.

Hauptmann's plays

In Hauptmann's wake several writers wrote in a straightforward naturalist manner. The best known among them were Hermann Sudermann, notable for *Ehre* (1889; "Honour"; Eng. trans. *What Money Cannot Buy*) and *Heimat* (1893; "Home"; Eng. trans. *Magda*), plays criticizing middle-class morality; and Max Halbe, whose *Jugend* (1893; "Adolescence") was a drama of young love.

In the course of the 19th century, German literature had increasingly abandoned an idealistic conception of man and turned to a more down-to-earth and deprecating appraisal of reality, reflecting the rise of positivist and materialist thought in science. This proved too narrow; and in consonance with the new relativist scientific cosmology of the 20th century, artistic imagination began to portray a more complex vision of the world.

The 20th century

Through the 20th century, German literature has reflected the social, political, and spiritual uncertainty of its surroundings. Early dissatisfaction with conventional literary forms led to experiments with new ones in an attempt to avoid sterility and to revitalize the language—aims that have emerged as dominant forces in modern literature.

MAJOR LITERARY TRENDS AND CONDITIONS

Impressionism. Impressionism evokes a mood or state of mind by emphasizing the impression made by an object on its observer. The poet Detlev von Liliencron provided an early example of this, as did Richard Dehmel. Writers influenced by Symbolism also had elements of Impressionism in their work. A successor of French Symbolism who had considerable effect on other writers was Stefan George, whose solemn, carefully composed verse aimed at asserting the lofty stature of poetry, which, for him, had a religious character. George founded the journal *Blätter für die Kunst* (1892; "Journal for Art") to publish his followers' poetry. Among those attracted by his work were the critic Friedrich Gundolf and the poet Karl Wolfskehl.

George's Symbolist poetry

Hugo von Hofmannsthal, an Austrian whose Impressionistic elements had their roots in Romanticism, declined to join George's circle. In his melodious poetry he delicately analyzed his sensibilities and was haunted by his obsession with the inadequacy of language completely to convey feeling. An essay, the "Chandos-Brief" (1902; "Letter by Lord Chandos"), records this sense of the inadequacy of words. The plays of the 1890s concerned the aesthete faced with the reality of this inadequacy. Later dramas, such as *Jedermann* (published 1911), an adaptation of *Everyman*, and *Das Salzburger grosse Welttheater* (1922; *The Great Salzburg Theatre of the World*), were religious in tone and borrowed from Baroque and medieval drama; a comedy, *Der Schwierige* (1921; *The Difficult Man*), analyzed a sophisticated mind inhibited by the weight of social tradition. His greatest public successes were his librettos, such as *Der Rosenkavalier* (1911), which Richard Strauss set to

music. Conscious of the heritage of European culture and of ethical responsibility, Hofmannsthal conveyed a strong awareness of moral issues in his work.

Vienna was now a major cultural centre. Arthur Schnitzler depicted its pre-1914 decadence. Karl Kraus relentlessly exposed the Viennese press and morals in his one-man satiric newspaper *Die Fackel* ("The Torch") and attacked World War I in his monumental drama *Die letzten Tage der Menschheit* (1915–17; "The Last Days of Mankind"). A cooler analyst was Robert Musil. His novel *Der Mann ohne Eigenschaften* (1930–43; *The Man Without Qualities*), one of the masterpieces of the age, ironically dissects modern incertitude, sham values, and political folly. Hermann Broch, too, gave a profound historical analysis in his trilogy *Die Schlafwandler* (1931–32; *The Sleepwalkers*). Joseph Roth was still another who depicted the decline of Austria-Hungary, as, for example, in the novel *Radetzkymarsch* (1932; *Radetzky March*).

The end-of-the-century analyses of decadence

Prague had also become a centre of writing: Franz Kafka, Rainer Maria Rilke, Franz Werfel, Max Brod, and Gustav Meyrink wrote and lived in Prague. Their work inclined toward the esoteric and was strongly influenced by Symbolism. One last generation of Prague writers, which included Hermann Grab, Johannes Urzidil, and Franz Wurm, became active in exile during and after World War II.

Symbolism. Like most poets of the time, Rilke was indebted to Symbolism. His melodious verse, which made him one of Germany's great lyric poets, gave his vision of reality compelling power. This was clearly evident in *Das Stunden-Buch* (1905; "The Book of Hours"), which described a search for spiritual health in a hostile urban civilization. Rilke's conception of the universe, God, and death was determined by a quest for artistic fulfillment. In *Neue Gedichte* (1907–08; "New Poems") he ceased to depict his subjective response to spiritual isolation and instead attempted to give a more objective view of life, art, and nature. In *Die Aufzeichnungen des Malte Laurids Brigge* (1910; *The Notebooks of Malte Laurids Brigge*) he studied the disintegration of an artistic sensibility alienated by the modern world. The *Duineser Elegien* (1923; *Duino Elegies*) summed up his spiritual struggles in complicated severe verse, while *Die Sonette an Orpheus* (1923; *Sonnets to Orpheus*) sought to show poetry's power to transmute problems of existence and to justify reality.

Rilke's view of God

Symbolism also influenced prose writers, not least Thomas Mann, who attempted to use symbol and myth in narratives that started from a clinical analysis of modern man's mental and physical state. Mann's characterization was Impressionist, but impressions became leitmotivs conveying the power of the subconscious. His work was influenced by the philosophers Schopenhauer and Nietzsche; yet his portrayal of social change, of the impact of ideology, was an organic part of the story, and he was occupied, as always, with the status of the artist in society. In his early works art was symbolic of decadence, an overrefinement no longer acceptable; but with *Der Zauberberg* (1924; *The Magic Mountain*), Mann, like Rilke, emphasized the constructive qualities of art. His great novels examined different facets of his age: *Buddenbrooks* (1900) explored bourgeois society; *Der Zauberberg* dealt with intellectual corruption; *Doktor Faustus* (1947) examined the German mind and character during the Third Reich. His novellas—*e.g., Tonio Kröger* and *Tristan* (both 1903) and *Der Tod in Venedig* (1912; *Death in Venice*)—depicted the same themes. Irony and resultant ambiguity were characteristic of Mann. His increasingly complex style reflected the complexity of his mind and his study of the history of ideas. Mann's last work, *Bekenntnisse des Hochstaplers Felix Krull* (1922 and revised 1954; *The Confessions of Felix Krull, Confidence Man*), humorously stated the doubt that ran through all his work—whether the pursuit of beauty and, hence, pursuit of culture and art were not in the end a great deception.

Mann's analyses of his age

Hermann Hesse was influenced by Neoromanticism and concentrated on man's spiritual conflict. His novels, from his first success, *Peter Camenzind* (1904), portrayed the struggle of individuals in a world hostile to sensitivity or explored the subconscious and the balance between sensuality and the spirit. In *Der Steppenwolf* (1927), he examined the conflict between the bourgeois world and the sensitive outsider, here resolved by self-abandonment to fantasy; in *Das Glasperlenspiel* (1943; "The Glass-bead Game"; Eng. trans. *Magister Ludi*) he questions the whole purpose of civilization.

Ricarda Huch's novels, like those of Thomas Mann and his brother Heinrich, emphasized the individual's independence and dignity. Heinrich Mann savagely attacked social and political abuses in novels such as *Professor Unrat* (1905; *Small Town Tyrant,* or *The Blue Angel*) and *Der Untertan* (1918; *The Patrioteer*).

Expressionism. Expressionism was the key movement in German literature, as it was in painting, during and immediately after World War I. Expressionism emphasized the inner significance of things and not their external forms. Actually anticipating the war, it depicted the disintegration of the world and proclaimed a quest for the "New Man."

Frank Wedekind's dramas were forerunners of this style. For example, the plays constituting what is commonly known as the Lulu Tragedy, *Erdgeist* (1895; *Earth Spirit*) and *Die Büchse des Pandora* (1895; *Pandora's Box*), had pilloried bourgeois morality and broken with dramatic convention. The first fully Expressionist drama, however, was Johannes Reinhard Sorge's *Bettler* (1912; "The Beggar"), in which characters appeared as abstract functions in each other's lives. This play, like those of Walter Hasenclever, Paul Kornfeld, Fritz von Unruh, Ernst Barlach the sculptor, and Oskar Kokoschka the painter, was characterized by a quest for the essence of things, for the ideas behind personality and spiritual meaning in life. Ernst Toller wrote political plays employing an Expressionist technique in *Die Maschinenstürmer* (published 1922; *The Machine-Wreckers*). Georg Kaiser, the leading Expressionist playwright, moved from naturalism through Expressionism to a mature traditional style, while Carl Sternheim unmasked bourgeois pretensions by means of a shrill satire of contemporary language.

Expressionist poetry was equally nonreferential, attaining coherence through its associative power. The chief poets were Ernst Stadler, Georg Heym, Georg Trakl, August Stramm, Gottfried Benn, and Else Lasker-Schüler. The mainspring of Expressionist verse was a horror over urban life and over the collapse of civilization. Portrayals of this apocalyptic vision range from Trakl's moving lamentation to a macabre, cynical detachment in Benn's early verse. The vision of meaninglessness was extended in the so-called Dadaism of poets such as Jean (Hans) Arp and Yvan Goll. An absurd view of life was also featured in the nonsense verse of Christian Morgenstern (*Galgenlieder,* 1905; *Gallows Songs*).

Franz Kafka shared this negative vision. His parables, stories, and novels seem to epitomize the problems of modern life. With the stark clarity of a nightmare, he depicted the horror and uncertainty of human existence. In *Das Urteil* (1913; *The Sentence,* or *The Judgment*), a father condemns his son to death; in *Die Verwandlung* (1915; *The Transformation,* or *Metamorphosis*), a man turns into a beetle; and in *In der Strafkolonie* (1919; *In the Penal Settlement*), a torture machine runs wild. In his posthumously published novels, notably *Der Prozess* (1925; *The Trial*) and *Das Schloss* (1926; *The Castle*), the individual is trapped in a labyrinth of anxiety and guilt and crushed by unfathomable forces. There is also humour in Kafka, sometimes grotesque and at others sublime. The terror of his art, however, is only alleviated by its moral strength.

The nightmare vision of Franz Kafka

Kafka's themes recall Expressionism, but his classically balanced style echoes the prose of the Swiss Robert Walser, author of *Jakob von Gunten* (1929). The Expressionist novel is best exemplified by Alfred Döblin's *Berlin Alexanderplatz* (1929; "Alexander Square, Berlin"; Eng. trans. *The Story of Franz Biberkopf*).

Post-Expressionism and Social Realism. After 1918 Expressionism gave way to Social Realism through which writers hoped to gain objectivity. The first subjects chosen were World War I and its aftermath: Arnold Zweig's *Streit um den Sergeanten Grischa* (1927; *The Case of Sergeant Grischa*), Erich Maria Remarque's *Im Westen*

Nichts Neues (1929; *All Quiet on the Western Front*), and Hans Fallada's (Rudolf Ditzen's) and Erich Kästner's works documented the war and postwar society.

This new objectivity continued in the work of Anna Seghers and in Carl Zuckmayer's *Hauptmann von Köpenick* (1931; *The Captain of Köpenick*) and *Des Teufels General* (produced 1946; "The Devil's General"). Ernst Jünger explored the philosophical implications of technology and modern civilization; *Auf den Marmorklippen* (1939; *On the Marble Cliffs*), a criticism of the Third Reich, revealed his inability to conceive imaginatively individual character. The poetry and prose of Jünger's brother Friedrich exhibited greater lyrical and narrative skill.

The Third Reich disrupted the continuity of German literary life. Under the Nazis, talented writers either left Germany, were driven out, were forced into silence, or were exterminated, and good writing did not come to the surface until after World War II.

LITERATURE AFTER WORLD WAR II

The Nazi era, the war, and its aftermath profoundly affected the course of German literature after 1945. Writers felt impelled to come to terms with their historical and political situation. In the West the Group 47 brought together almost all important writers from Heinrich Böll to Günter Grass, and for more than 20 years it helped to shape the literary climate. In the East a new literature emerged from the dialogue with Marxism, but an increasing number of writers moved to the West.

The atmosphere after 1945, known as Zero Hour, was captured in the stories of Wolfgang Borchert, in Ilse Aichinger's novel *Die grössere Hoffnung* (1948; "The Greater Hope"), and in the early fiction of Heinrich Böll.

Böll's writings

Böll's humane analysis of postwar problems in such novels as *Haus ohne Hüter* (1954; *The Unguarded House*) made him the most representative German writer of the period. He was awarded the Nobel Prize for Literature for 1972. In his comic short stories and the novel *Ansichten eines Clowns* (1963; *The Clown*), Böll looked critically at modern life, and he did so increasingly in later works such as *Die verlorene Ehre der Katharina Blum* (1974; *The Lost Honor of Katharina Blum*) and *Fürsorgliche Belagerung* (1979; *The Safety Net*).

Grass's portrayal of modern German history

Günter Grass provided a more exuberant, amoral, and grotesque picture of German history in his Danzig trilogy: *Die Blechtrommel* (1959; *The Tin Drum*), *Katz und Maus* (1961; *Cat and Mouse*), and *Hundejahre* (1963; *Dog Years*). In *Das Treffen in Telgte* (1979; *The Meeting at Telgte*), he wrote a satiric epitaph on Group 47. Another work on an ambitious scale was Uwe Johnson's tetralogy *Jahrestage* (1970–83; *Anniversaries*), which gives a minute analysis of the German situation. Other notable contemporary German novelists include Wolfgang Koeppen, Alfred Andersch, Siegfried Lenz, and Martin Walser. Among women writers who established major reputations were the Austrian Ingeborg Bachmann, author of the novel *Malina* (1971), and the East German Christa Wolf, who wrote several stories and the novels *Der geteilte Himmel* (1963; *Divided Heaven*) and *Nachdenken über Christa T.* (1968; *The Quest for Christa T.*).

Several novelists continued the prewar Viennese tradition. George Saiko's *Auf dem Floss* (1948; "On the Raft") and Albert Paris Gütersloh's *Sonne und Mond* (1962; "Sun and Moon") provided brilliant symbols of the vanished Austrian world. Heimito von Doderer's complex but absorbing novels *Die Strudlhofstiege* (1951; "The Strudlhof Steps") and *Die Dämonen* (1956; *The Demons*) achieved a sustained and vivid panorama of prewar society. Elias Canetti's inspired novel about a world gone mad, *Die Blendung* (1935; "The Deception"; Eng. trans. *Auto-da-Fé*), won international recognition. His plays revolved around related themes, as did his sociological study *Masse und Macht* (1960; *Crowds and Power*). Canetti was awarded the Nobel Prize for Literature for 1981. His other publications include aphorisms, criticism, and a remarkable two-volume autobiography. Other novelists in the Austrian tradition include H.G. Adler, Peter von Tramin, Wolfgang Georg Fischer, and the comic writer Fritz von Hertzmannovsky-Orlando.

German poetry of the post-World War II period has been rich and varied. Two opposing trends, however, have dominated the scene: one toward difficult, esoteric verse, and the other toward a simple, plain style. Hans Carossa and other poets of the older generation continued to publish, and the war's victims such as Jesse Thoor, Gertrud Kolmar, and Albrecht Haushofer were discovered. Poets, like authors of prose fiction, caught the mood of Zero Hour. Later, Paul Celan developed increasingly intricate, hermetic verses; Johannes Bobrowski recreated myth; and Christine Lavant wrote symbolic yet folklike poems. The poet Nelly Sachs was awarded the Nobel Prize for Literature for 1966. Hans Magnus Enzensberger manufactured what is often called consumer poetry, and Wolf Biermann and Erich Fried popularized political verse. Later poets were more subjective.

The plays of Bertolt Brecht

The giant of European theatre was Bertolt Brecht, much of whose work had been completed almost 30 years before he received international acclaim. His first successful play, *Trommeln in der Nacht* (performed 1922; *Drums in the Night*), and the satiric opera *Die Dreigroschenoper* (1928, music by Kurt Weill; *The Threepenny Opera*) dated from the 1920s, when he was developing his theory of "epic theatre" and his social criticism, which led him to Marxism. During the Third Reich he went into exile, where he wrote his most mature plays: *Mutter Courage und ihre Kinder* (1941; *Mother Courage and Her Children*), *Leben des Galilei* (1943; *The Life of Galileo*), and *Der kaukasische Kreidekreis* (1948; *The Caucasian Chalk Circle*). Besides being a gifted playwright, Brecht was a highly influential theorist and a poet of considerable talent.

Other significant contributions to modern German drama during the postwar years have come from Switzerland. The two best-known contemporary Swiss writers, Max Frisch and Friedrich Dürrenmatt, both have made bold experiments with dramatic form. Dürrenmatt's *Besuch der alten Dame* (1955; *The Visit*) and *Die Physiker* (1961; *The Physicists*) and Frisch's *Nun singen sie wieder* (1946; "Now Sing Again") and *Andorra* (1961) are modern morality plays. The latter's novels, *Stiller* (1954), *Homo Faber* (1957), and *Der Mensch erscheint im Holozän* (1979; *Man in the Holocene*), are investigations into the place of the intellectual in the modern world. Both writers criticize the emotional sterility of modern life.

Documentary school of drama

Several German dramatists received international attention, notably the politically oriented so-called documentary school represented by Rolf Hochhuth's *Der Stellvertreter* (1963; *The Deputy*) and *Die Soldaten* (1967; "The Soldiers") and Peter Weiss's *Die Ermittlung* (1965; *The Investigation*). Weiss's *Die Verfolgung und Ermordung des Jean-Paul Marats . . .* (1964; *The Persecution and Assassination of Jean-Paul Marat . . .*; often referred to simply as *Marat/Sade*) was a political dramatic spectacle. His *Hölderlin* (1971) reinterpreted the poet's life. Other German dramatists were Tankred Dorst, Botho Strauss, and Franz Xaver Kroetz. The Austrian dramatists Wolfgang Bauer and Peter Handke excelled at provoking their audiences. Handke's plays were *Publikumsbeschimpfung* (1966; "Abusing the Audience"), *Kaspar* (1968), and *Der Ritt über den Bodensee* (1971; *The Ride Across Lake Constance*). Handke turned increasingly to prose, as in *Die Angst des Tormanns beim Elfmeter* (1970; *The Goalie's Anxiety at the Penalty Kick*), the semiautobiographical *Wunschloses Unglück* (1972; *A Sorrow Beyond Dreams*), and *Die linkshändige Frau* (1976; *The Left-handed Woman*), and he popularized the techniques of experimental writing.

The New German Radio Play

Much of the writing produced since the late 1940s has been decidedly innovative and experimental. The poet Günther Eich perfected the then new genre of the radio play in *Träume* (1951; "Dreams"). Many others, including Bachmann, Böll, and Dürrenmatt, tried their hand at the genre as well. It was followed by the New German Radio Play, in which meaning gave way to experimentation with form and sound, as in Franz Mon's *ich bin der ich bin die* (broadcast 1971; "i am he i am she"). This genre emerged from concrete poetry, a verse style developed by Eugen Gomringer and the Vienna Group of Hans Carl Artmann, Gerhard Rühm, and Konrad Bayer, as well as by Ernst

Jandl and Friederike Mayröcker. Such poets treated language as a physical object, both visually and acoustically. Bayer, and Helmut Heissenbüttel in *D'Alembert's Ende* (1970; "D'Alembert's Death"), extended the style into the genre of the novel, as did Jürgen Becker. The doyen of experimental prose, however, was Arno Schmidt, whose works include *Die Gelehrtenrepublik* (1957; "The Republic of Letters"; Eng. trans. *The Egghead Republic* or *Republica Intelligentsia*); his magnum opus, *Zettels Traum* (1970; "Zettel's Dream"), developed complex linguistic play within a visual structure.

Trends in literary criticism

Literary criticism also became more complex during the 20th century. It increasingly supplanted 19th-century positivism with a variety of approaches that treated literature in the light of the history of ideas as defined by Wilhelm Dilthey. Two very different exponents of this method were Fritz Strich and Hermann August Korff. After World War II stylistic analysis, which made the work of literature rather than the writer the focus of inquiry, emerged as the most popular tendency. Its leading critic was for many years Emil Staiger. This text-oriented school, however, underwent a crisis in the 1960s, which coincided with the discovery of the prewar esoteric Marxist criticism of Walter Benjamin. Two new trends emerged: a sociopolitical approach and a more self-conscious, theoretical one. The ensuing debate over method stimulated a wide variety of critical styles, reflecting the prevailing relativism of the age.

The end of the century

Günter Grass and Christa Wolf continued at the forefront of German writers. Grass, influenced by the multiple-text bias of what has come to be called postmodernism, produced *Der Butt* (1977; *The Flounder*) and *Die Rättin* (1986; *The Rat*), two works composed of varying types of narrative. Among his later novels was *Ein weites Feld* (1995; *Too Far Afield*), a novel set in Berlin that deals in part with the subject of Germany's reunification. Reunification also influenced Wolf's *Was bleibt* (1990; *What Remains*), about the author's surveillance by East German police, and the narrator in Thomas Hettche's *Nox* (1995; [Latin: "At Night"]) has his throat slit the night the Berlin Wall comes down. East Germany itself is targeted in Reiner Kunze's *Deckname "Lyrik"* (1990; "Code Name 'Lyric'"). Kunze, like Wolf, had been under the surveillance of the East German secret police, who are also featured in Thomas Brussig's satiric novel *Helden wie wir* (1995; *Heroes Like Us*).

The country's Nazi past remained an important topic. Monika Maron's *Stille Zeile Sechs* (1991; *Silent Close No. Six*) deals with postwar guilt and also with East Germany's communist experience. Marcel Beyer's *Flughunde* (1995; "Flying Foxes," Eng. trans. *Flughunde*) centres on the death of Nazi propagandist Joseph Goebbels' children. In Thomas Lehr's *Frühling* (2001; "Spring"), the suicidal narrator has a father who had served as a doctor in a concentration camp. *Vati* (1987; "Daddy"), by Peter Schneider, presents another Nazi father, based on the infamous doctor Josef Mengele. Thomas Bernhard's *Ein Zerfall* (1986; *Extinction*) deals in part with Austria's collaboration with the Nazis. W.G. Sebald's important novel *Austerlitz* (2001; Eng. trans. *Austerlitz*) features a character who explores the past he had escaped by being adopted by an English couple. Another significant work along these lines is Christoph Ransmayr's *Morbus Kitahara* (1995; *The Dog King*), set in an imaginary Germany returned to preindustrialization as a punishment.

At the turn of the century, then, Germany's writers continued to be haunted by the country's recent past, on the one hand, and exercised by the possibilities and the problems of reunification, on the other. The residual sense of crime and guilt for the Nazi years, together with a more general Western trend toward exploring the outer limits of individual behaviour, is epitomized in works such as Patrick Süskind's *Das Parfum: Die Geschichte eines Mörders* (1985; *Perfume: The Story of a Murderer*) and in his Kafkaesque *Die Taube* (1987; *The Pigeon*); and the profusion of voices resulting from the complicated state of German life is well represented by Günter Grass's *Mein Jahrhundert* (1999; *My Century*), with its various texts—history, memoir, and, to encompass the tragic and fantastic aspects of Germany's past, fiction.

BIBLIOGRAPHY. Good general approaches include WERNER P. FRIEDERICH, *History of German Literature*, 2nd ed. (1961); WOLFGANG BEUTIN *et al., A History of German Literature: From the Beginnings to the Present Day*, 4th ed. (1993; originally published in German, 1979); and HELEN WATANABE-O'KELLY (ed.), *The Cambridge History of German Literature* (1997).

Studies that focus on specific periods or trends of German literary history include J. KNIGHT BOSTOCK, *A Handbook on Old High German Literature*, 2nd ed. rev. by K.C. KING and D.R. MCLINTOCK (1976), an essential reference work; FRANZ H. BÄUML, *Medieval Civilization in Germany, 800–1273* (1969); JOACHIM BUMKE, *Courtly Culture: Literature and Society in the High Middle Ages* (1991, reissued 2000; originally published in German, 1986); J. HUIZINGA, *The Waning of the Middle Ages: A Story of the Forms of Life, Thought, and Art in France and The Netherlands in the XIVth and XVth Centuries*, trans. by F. HOPMAN (1924, reissued 2001; originally published in Dutch, 1919), a classic study of late medieval decadence; ARCHER TAYLOR, *Problems in German Literary History of the Fifteenth and Sixteenth Centuries* (1939, reprinted 1966); ROY PASCAL and HANNAH PRIEBSCH CLOSS, *German Literature in the Sixteenth and Seventeenth Centuries* (1968, reprinted 1979), one of the best English introductions to this period; ROBERT M. BROWNING, *German Baroque Poetry, 1618–1723* (1971); FRIEDHELM RADANT, *From Baroque to Storm and Stress, 1720–1775* (1977), a most useful survey; WALTER H. BRUFORD, *Germany in the Eighteenth Century: The Social Background of the Literary Revival* (1935, reprinted 1971), and *Culture and Society in Classical Weimar, 1775–1806* (1962), informative studies of the cultural milieu of the period; HENRY B. GARLAND, *Storm and Stress* (1952), an adequate introduction; WILLIAM D. ROBSON-SCOTT, *The Literary Background of the Gothic Revival in Germany: A Chapter in the History of Taste* (1965), on art as well as literature; T.J. REED, *The Classical Centre: Goethe and Weimar, 1772–1832* (1980, reissued 1986), a masterly study; ERNST L. STAHL and W.E. YUILL, *German Literature of the Eighteenth and Nineteenth Centuries* (1970), a solid introductory study; ROBERT C. HOLUB, *Reflections of Realism: Paradox, Norm, and Ideology in Nineteenth-Century German Prose* (1991); GLYN TEGAI HUGHES, *Romantic German Literature* (1979), an excellent survey; GEORG BRANDES, *Main Currents in 19th-Century Literature*, vol. 6, *The Young Germany* (1901–05, reissued 1975; originally published in Danish, 1872–90), a thorough work on European comparative literature; RONALD GRAY, *The German Tradition in Literature, 1871–1945* (1965), an interesting study; JETHRO BITHELL, *Modern German Literature, 1880–1950*, 3rd ed. (1959, reprinted 1963), a useful reference work; AUGUST CLOSS (ed.), *Twentieth Century German Literature* (1969), an excellent introduction, one in the series of Introductions to German Literature; RAYMOND FURNESS, *The Twentieth Century, 1890–1945* (1978); WALTER H. SOKEL, *The Writer in Extremis: Expressionism in Twentieth-Century German Literature* (1959, reissued 1964); RUSSELL A. BERMAN, *The Rise of the Modern German Novel: Crisis and Charisma* (1986); STEPHEN BROCKMANN, *Literature and German Reunification* (1999); and ERNESTINE SCHLANT, *The Language of Silence: West German Literature and the Holocaust* (1999).

(W.W.C./D.G.D./A.Gi./H.S.R./J.D.Ad./J.N./Ed.)

Germany

The core area of the German-speaking peoples, the Federal Republic of Germany (Bundesrepublik Deutschland) occupies a section of north-central Europe traversing the continent's main physical divisions, from the outer ranges of the Alps, northward across the varied country of the Central German Uplands, and then across the North German Plain, or Lowlands. It is bounded at its extreme north on the Jutland Peninsula by Denmark. East and west of the peninsula, the Baltic Sea (Ostsee) and North Sea coasts, respectively, complete the northern border. To the west, Germany borders on The Netherlands, Belgium, and Luxembourg; to the southwest, on France. It shares its entire southern boundary with Switzerland and Austria. In the southeast, the border with the Czech Republic corresponds to an earlier boundary of 1918, renewed by treaty in 1945. The easternmost frontier adjoins Poland along the northward course of the Neisse River and subsequently the Oder to the Baltic Sea, with a westward deviation in the north to exclude the former German port city of Stettin (Polish: Szczecin) and the Oder mouth. This border reflects the loss of Germany's eastern territories mandated in the Potsdam agreement among the victorious World War II Allies and reaffirmed by subsequent governments.

Of historical, if no longer political, importance is the long internal boundary that for 40 years partitioned Germany into two nations. It was based on the lines of demarcation agreed upon at the Yalta Conference of 1945, separating the then-Soviet occupation zone of Germany from the zones occupied by the Western allies, on which territory West Germany subsequently emerged. On the East German side, this boundary was, until the fall of the communist government in 1989, marked by defenses in depth designed to prevent the escape of the population. The 185 square miles (480 square kilometres) of the "island" of West Berlin were similarly ringed from 1961 to 1989 by the Berlin Wall running through the city and a heavily guarded wire-mesh fence in the areas abutting the East German countryside. The city declined in national and international significance until the events of 1989–90 restored a united Berlin as the capital of a united Germany. In June 1991 the Bundestag voted to move the seat of government from Bonn to Berlin.

The republic has a maximum north-south extent of about 520 miles (840 kilometres), between latitudes 47° and 55° N, and an east-west maximum (across the middle of the country) of about 385 miles (620 kilometres), between longitudes 6° and 15° E. It has an area of 137,735 square miles (356,733 square kilometres).

The constitution of the republic, adopted in 1949, gives most government powers to its constituent *Länder* (states), which at the time of unification numbered 11 (including West Berlin, which, however, had the special status of a *Land* without voting rights) but which grew to 16 (including the reunited Berlin) upon the accession of East Germany; only matters of undoubted importance for the nation as a whole, such as defense and foreign affairs, are reserved to the federal government. At both the state and federal levels, parliamentary democracy prevails. The Federal Republic has been a member of the North Atlantic Treaty Organization (NATO) since 1955 and was a founding member of the European Economic Community in 1957. During the four decades of partition, the Federal Republic concluded a number of agreements with the U.S.S.R. and East Germany, which it supported to some extent economically in return for various concessions with regard to access to Berlin and humanitarian matters. West Germany's rapid economic recovery in the 1950s (*Wirtschaftswunder,* or "economic miracle") brought it into a leading position among the world's economic powers, a position that it has subsequently maintained. Reunion with the eastern territories, while constituting a huge financial burden at the outset, was expected eventually to strengthen this standing.

The constituent states, including those delineated in the former eastern sector in 1990, follow the historic lines of the political divisions of the former Reich. They enjoy considerable political autonomy within the federative structure, especially in such areas as education, finance, and law enforcement. Each has its own equivalent of a prime minister, a parliament or diet, and provincial ministries. Their strong voice in the upper levels of the federal government is unique among Western republics.

The largest of the states is Bavaria (Bayern), the richest is Baden-Württemberg, and the most populous is North Rhine-Westphalia (Nordrhein-Westfalen). In the extreme north lies Schleswig-Holstein, south of which is Lower Saxony (Niedersachsen). The ancient free Hanseatic cities of Hamburg and Bremen rank as states in their own right. Rhineland-Palatinate (Rheinland-Pfalz) and Hesse (Hessen) cover the central area of the republic; the Saarland is a pocket in the southwestern corner of the Pfälzer Mountains (Pfälzerwald). Five states, in addition to that of Berlin, were formed in 1990 from the territory of the former Democratic Republic: Mecklenburg-West Pomerania (Mecklenburg-Vorpommern) in the north; Brandenburg (surrounding Berlin) in the east; Saxony (Sachsen) south of Brandenburg and bordering on the northwestern Czech Republic; Thuringia (Thüringen) between Saxony and Hesse; and Saxony-Anhalt (Sachsen-Anhalt) between Brandenburg and Lower Saxony. They are small in comparison with most of the states of the original Federal Republic.

Historically, the German peoples have been characterized by almost perpetual political disunity and fluctuating boundaries in their central European position. The Treaty of Versailles, at the end of World War I, established a string of states, notably Poland and Czechoslovakia, at the expense of former Reich or Austrian territory, containing substantial German-speaking minorities. Germany's defeat in 1945 was even more devastating for the German-speaking peoples. The Soviet Union and Poland occupied nearly a quarter of Germany's former territory, and they proceeded to expel most of the German-speaking populations. At the same time, German minorities were expelled from some of the post-Versailles successor states, notably Czechoslovakia. In a brutal fashion, therefore, the frontiers of post-1945 Germany came to be largely coincident with the distribution of German-speaking people; only Austria, Liechtenstein, and German-speaking Switzerland remained outside. (T.H.El./G.H.K.)

This article first discusses the physical and human geography of Germany. Following that section is a discussion of the history of the Germans from ancient times to the present, including the reunification of the German state. For detailed coverage of the city of Berlin, see the article BERLIN.

The article is divided into the following sections:

Physical and human geography 40
The land 40
Relief
Drainage and soils
Climate
Plant and animal life
Settlement patterns
The people 48
Ethnic structures
Linguistic diversity
Religions
Demographic trends

Population structure
The economy 51
From partition to reunification
Economic unification and beyond
Agriculture, forestry, and fishing
Resources and power
Manufacturing
Finance
Trade
Services
Labour and taxation
Transportation and telecommunications
Administration and social conditions 57
Government
Justice
Political institutions
Armed forces
Law enforcement
Education
Health and welfare
Housing
Standards of living
Cultural life 63
The cultural milieu
State of the arts
Cultural institutions
History 67
Ancient history 67
Coexistence with Rome to AD 350
The migration period
Merovingians and Carolingians 69
Merovingian Germany
The rise of the Carolingians and Boniface
Charlemagne
The emergence of Germany
Germany from 911 to 1250 70
The 10th and 11th centuries
Germany and the Hohenstaufen, 1125–1250
Germany from 1250 to 1493 78
1250 to 1378
1378 to 1493
Germany from 1493 to *c.* 1760 87
Reform and Reformation: 1493–1555
The confessional age: 1555–1648
Territorial states in the age of absolutism
Germany from *c.* 1760 to 1871 98
Germany to 1815
The age of Metternich and the era of unification: 1815–71
Germany from 1871 to 1945 111
The German Empire, 1871–1918
The German Republic, 1918–33
The Third Reich, 1933–45
The era of partition 124
Allied occupation and the formation of the two Germanies, 1945–49
Political consolidation and economic growth, 1949–69
Ostpolitik and reconciliation, 1969–89
The reunification of Germany 129
Bibliography 130

PHYSICAL AND HUMAN GEOGRAPHY

The land

The major lineaments of Germany's physical geography are not unique. The country spans the great east-west morphological zones that are characteristic of the western part of central Europe. In the south, Germany impinges on the outermost ranges of the Alps. From there it extends across the Alpine Foreland (Alpenvorland), the plain on the northern edge of the Alps. Forming the core of the country is the large zone of the Central German Uplands, which is part of a wider European arc of territory stretching from the Massif Central of France in the west into the Czech Republic, Slovakia, and Poland in the east. In Germany it manifests itself as a landscape with a complex mixture of forested block mountains, intermediate plateaus with scarped edges, and lowland basins. In the northern part of the country, the North German Plain, or Lowland, forms part of the greater North European Plain, which broadens from the Low Countries eastward across Germany and Poland into Belarus, the Baltic States, and Russia and extends northward through Schleswig-Holstein into the Jutland peninsula of Denmark. The North German Plain is fringed by marshes, mud flats, and the islands of the North and Baltic seas. In general, Germany has a south-to-north drop in elevation, from a maximum 9,718 feet (2,962 metres) in the Zugspitze of the Bavarian Alps to a few areas slightly below sea level in the north.

It is a common assumption that surface configuration reflects the underlying rock type; a hard resistant rock such as granite will stand out, whereas a softer rock such as clay will be weathered away. However, this assumption is not always borne out. The Zugspitze, for example, is Germany's highest summit not because it is composed of particularly resistant rocks but because it was raised by the mighty earth movements that began in the Tertiary Period, some 37 to 24 million years ago and created the Alps, Europe's highest and youngest fold mountains. Another powerful force determining surface configuration is erosion, mainly by rivers. In the Late Carboniferous Epoch, (360 to 286 million years ago), an earlier mountain chain, the Hercynian, or Variscan, mountains, had crossed Europe in the area of the Central German Uplands. Yet the forces of erosion were sufficient to reduce these mountains to almost level surfaces, on which a series of secondary sedimentary rocks (those of Permian to Jurassic age [286 to 144 million years old]) were deposited. The entire formation was subsequently fractured and warped under the impact of the Alpine orogeny. This process was accompanied by some volcanic activity, which left behind not only peaks but also a substantial number of hot and mineral springs. Torrential erosion occurred as the Alpine chains were rising, filling the furrow that now constitutes the Alpine Foreland. The pattern of valleys eroded by streams and rivers has largely given rise to the details of the present landscape. Valley glaciers emerging from the Alps and ice sheets from Scandinavia had some erosive effect, but they mainly contributed sheets of glacial deposits. Slopes outside the area of the actual ice sheets—those under tundra conditions and unprotected by vegetation—were rendered less steep by the periglacial slumping of surface deposits under the influence of gravitation. Winds blowing over unprotected surfaces fringing the ice sheets picked up fine material known as loess; once deposited, it became one of Germany's outstanding soil-parent materials. Coarser weathered material was carried into alluvial cones and gravel-covered river terraces, as in the Rhine Rift Valley (Rhine Graben).

Forces of erosion

The detailed morphology of Germany is significant in providing local modifications to climate, hydrology, and soils.

RELIEF

The Central German Uplands. Geographically the Central German Uplands form a region of great complexity. Under the impact of the Alpine orogeny, the planed-off remnants of the former Hercynian mountains were shattered and portions thrust upward to form block mountains, with sedimentary rocks preserved between them in lowlands and plateaus. The Central German Uplands may be divided into three main parts: a predominantly lowland country in the south, an arc of massifs and plateaus running from the Rhenish Uplands to Bohemia, and a fairly narrow northern fringe, composed of folded secondary rocks.

Southern Germany. In southern Germany, Hercynian massifs are of restricted extent. The Black Forest (Schwarzwald) must once have been continuous with the Vosges Mountains in what is now France, but they were broken apart through the sinking of a central strip to form the Rhine Rift Valley, which extends 185 miles (300 kilometres) in length. The Black Forest reaches its greatest elevation in the south (Mount Feld, 4,898 feet [1,493 metres]) and declines northward beneath secondary sedi-


Scale 1:3,920,000
1 inch equals approx. 62 miles
Conic Projection
NORTH SEA
BALTIC SEA
DENMARK
NORTH FRISIAN ISLANDS
SYLT
HELGOLAND
EAST FRISIAN ISLANDS
Wadden Sea
SCHLESWIG-HOLSTEIN
Kiel
Hamburg
Bremen
Bremerhaven
LÜNEBURG HEATH
MECKLENBURG-WEST POMERANIA
RÜGEN
Rostock
Schwerin
Lübeck
Szczecin
UCKERMARK
PRIGNITZ
POLAND
Berlin
BERLIN
BRANDENBURG
Potsdam
ALTMARK
SAXONY-ANHALT
Magdeburg
Hannover
Wolfsburg
Braunschweig
Hildesheim
Salzgitter
Osnabrück
Münster
Bielefeld
Paderborn
Göttingen
Kassel
Dortmund
Essen
Duisburg
Düsseldorf
Cologne
Bonn
Koblenz
Frankfurt am Main
Wiesbaden
Mainz
Darmstadt
Mannheim
Heidelberg
Stuttgart
Ulm
Augsburg
Munich
Leipzig
Dresden
Chemnitz
Zwickau
Erfurt
Jena
Gera
Dessau
Cottbus
Prague
CZECH REPUBLIC
AUSTRIA
SWITZERLAND
LIECHTENSTEIN
FRANCE
BELGIUM
LUXEMBOURG
THE NETHERLANDS
EMSLAND
MÜNSTERLAND
NORTH RHINE-WESTPHALIA
TEUTOBURG FOREST
HARZ MTS.
FLÄMING HILLS
LAUSITZ
OBERLAUSITZ
SAXONY
THURINGIA
HESSE
RHINELAND-PALATINATE
EIFEL UPLAND
HUNSRÜCK UPLAND
SAARLAND
WESTER FOREST
TAUNUS MTS.
RHÖN MTS.
SPESSART
ODEN FOREST
HAARDT MTS.
BADEN-WÜRTTEMBERG
ORTENAU
BLACK FOREST
SWABIAN JURA
FRANCONIAN JURA
BOHEMIAN FOREST
BAVARIAN FOREST
BAVARIA
BAVARIAN ALPS
ALPS
ORE MOUNTAINS
Cities over 1,000,000
Cities 500,000 to 1,000,000
Cities 100,000 to 500,000
Cities under 100,000
National capitals
Political subdivision capitals
HAMBURG Political subdivision names
International boundaries
Political subdivision boundaries
Canals
Dams
Swamps and marshes
National parks
Spot elevations in metres (1 m = 3.28 ft)
© Encyclopaedia Britannica Inc.

MAP INDEX

Political subdivisions

Baden-Württemberg .... 48 30 N 9 00 E
Bavaria (Bayern) .... 49 00 N 11 30 E
Berlin .... 52 30 N 13 22 E
Brandenburg .... 52 00 N 13 30 E
Bremen .... 53 05 N 8 50 E
Hamburg .... 53 35 N 10 00 E
Hesse (Hessen) .... 50 30 N 9 15 E
Lower Saxony (Niedersachsen) .... 52 00 N 10 00 E
Mecklenburg-West Pomerania (Mecklenburg-Vorpommern) .... 53 45 N 13 00 E
North Rhine-Westphalia (Nordrhein-Westfalen) .... 51 30 N 7 00 E
Rhineland-Palatinate (Rheinland-Pfalz) .... 49 40 N 7 30 E
Saarland .... 49 15 N 7 00 E
Saxony (Sachsen) .... 51 00 N 13 30 E
Saxony-Anhalt (Sachsen-Anhalt) .... 52 00 N 11 40 E
Schleswig-Holstein .... 54 00 N 10 30 E
Thuringia (Thüringen) .... 51 00 N 11 00 E

Cities and towns

Aachen .... 50 46 N 6 06 E
Aalen .... 48 50 N 10 06 E
Achim .... 53 02 N 9 01 E
Ahlen .... 51 45 N 7 55 E
Altenburg .... 50 59 N 12 27 E
Amberg .... 49 27 N 11 52 E
Amorbach .... 49 39 N 9 14 E
Andernach .... 50 26 N 7 24 E
Anerbach-Buchholz .... 50 34 N 13 00 E
Ansbach .... 49 18 N 10 35 E
Apolda .... 51 01 N 11 30 E
Arnsberg .... 51 23 N 8 05 E
Arnstadt .... 50 50 N 10 57 E
Aschaffenburg .... 49 59 N 9 09 E
Aschersleben .... 51 45 N 11 28 E
Aue .... 50 35 N 12 42 E
Augsburg .... 48 22 N 10 53 E
Aurich .... 53 28 N 7 29 E
Bad Ems .... 50 20 N 7 43 E
Bad Gandersheim .... 51 52 N 10 02 E
Bad Harzburg .... 51 53 N 10 34 E
Bad Hersfeld .... 50 52 N 9 42 E
Bad Homburg [vor der Höhe] .... 50 13 N 8 37 E
Bad Kissingen .... 50 12 N 10 05 E
Bad Kreuznach .... 49 50 N 7 52 E
Bad Mergentheim .... 49 29 N 9 46 E
Bad Reichenhall .... 47 44 N 12 53 E
Bad Salzuflen .... 52 05 N 8 46 E
Baden-Baden .... 48 45 N 8 15 E
Bamberg .... 49 52 N 10 52 E
Bautzen .... 51 11 N 14 26 E
Bayreuth .... 49 57 N 11 35 E
Berchtesgaden .... 47 38 N 13 00 E
Bergheim .... 50 58 N 6 39 E
Bergisch Gladbach .... 50 59 N 7 08 E
Berlin .... 52 30 N 13 22 E
Bernburg .... 51 48 N 11 44 E
Bernkastel-Kues .... 49 55 N 7 04 E
Bielefeld .... 52 02 N 8 32 E
Bingen .... 49 58 N 7 54 E
Blankenburg .... 51 47 N 10 57 E
Bocholt .... 51 50 N 6 36 E
Bochum .... 51 29 N 7 13 E
Bonn .... 50 44 N 7 06 E
Borna .... 51 07 N 12 30 E
Brandenburg .... 52 25 N 12 33 E
Braunschweig (Brunswick) .... 52 16 N 10 32 E
Breisach .... 48 02 N 7 35 E
Bremen .... 53 05 N 8 48 E
Bremerhaven .... 53 33 N 8 35 E
Brilon .... 51 24 N 8 35 E
Bruchsal .... 49 08 N 8 36 E
Brühl .... 50 50 N 6 54 E
Brunswick, see Braunschweig
Burg .... 52 16 N 11 51 E
Buxtehude .... 53 27 N 9 42 E
Calw .... 48 43 N 8 44 E
Celle .... 52 37 N 10 05 E
Chemnitz (Karl-Marx-Stadt) .... 50 50 N 12 55 E
Cloppenburg .... 52 51 N 8 02 E
Coburg .... 50 15 N 10 58 E
Cochem .... 50 08 N 7 09 E
Cologne (Köln) .... 50 56 N 6 57 E
Coswig .... 51 08 N 13 35 E
Cottbus .... 51 46 N 14 20 E
Crailsheim .... 49 09 N 10 05 E
Crimmitschau .... 50 49 N 12 23 E
Cuxhaven .... 53 53 N 8 42 E
Dachau .... 48 16 N 11 26 E
Darmstadt .... 49 52 N 8 39 E
Deggendorf .... 48 50 N 12 58 E
Delitzsch .... 51 32 N 12 21 E
Delmenhorst .... 53 03 N 8 37 E
Dessau .... 51 50 N 12 15 E
Detmold .... 51 56 N 8 53 E
Dinkelsbühl .... 49 04 N 10 19 E
Döbeln .... 51 07 N 13 07 E
Donauwörth .... 48 42 N 10 48 E
Dormagen .... 51 06 N 6 50 E
Dorsten .... 51 40 N 6 58 E
Dortmund .... 51 31 N 7 27 E
Dresden .... 51 03 N 13 45 E
Duisburg .... 51 26 N 6 45 E
Düren .... 50 48 N 6 29 E
Düsseldorf .... 51 13 N 6 46 E
Eberswalde-Finow .... 52 50 N 13 47 E
Eichstätt .... 48 53 N 11 11 E
Ellenberg .... 51 28 N 12 27 E
Eisenach .... 50 59 N 10 19 E
Eisenberg .... 50 58 N 11 54 E
Eisenhüttenstadt .... 52 09 N 14 39 E
Eisleben .... 51 32 N 11 33 E
Ellwangen .... 48 57 N 10 08 E
Elmshorn .... 53 45 N 9 39 E
Emden .... 53 22 N 7 13 E
Erding .... 48 18 N 11 56 E
Erfurt .... 50 59 N 11 02 E
Erlangen .... 49 36 N 11 01 E
Eschwege .... 51 11 N 10 04 E
Eschweiler .... 50 49 N 6 15 E
Essen .... 51 27 N 7 01 E
Esslingen .... 48 45 N 9 18 E
Ettlingen .... 48 57 N 8 24 E
Eutin .... 54 08 N 10 37 E
Falkensee .... 52 34 N 13 05 E
Flensburg .... 54 47 N 9 26 E
Forchheim .... 49 43 N 11 04 E
Forst .... 51 44 N 14 38 E
Frankenthal .... 49 32 N 8 21 E
Frankfurt .... 52 21 N 14 33 E
Frankfurt am Main .... 50 07 N 8 41 E
Freiberg .... 50 55 N 13 22 E
Freiburg [im Breisgau] .... 48 00 N 7 51 E
Freising .... 48 24 N 11 44 E
Freital .... 51 01 N 13 39 E
Friedenstadt .... 48 20 N 9 35 E
Friedberg .... 48 21 N 10 59 E
Friedrichshafen .... 47 39 N 9 29 E
Fulda .... 50 33 N 9 40 E
Fürstenfeldbruck .... 48 11 N 11 15 E
Fürstenwalde .... 52 22 N 14 04 E
Fürth .... 49 28 N 11 00 E
Furtwangen .... 48 03 N 8 12 E
Füssen .... 47 34 N 10 42 E
Ganderkesee .... 53 02 N 8 32 E
Garbsen .... 52 25 N 9 36 E
Garmisch-Partenkirchen .... 47 30 N 11 06 E
Gelsenkirchen .... 51 31 N 7 06 E
Genthin .... 52 24 N 12 10 E
Gera .... 50 52 N 12 05 E
Germering .... 48 08 N 11 22 E
Giessen .... 50 35 N 8 39 E
Glauchau .... 50 49 N 12 32 E
Glückstadt .... 53 47 N 9 25 E
Göppingen .... 48 42 N 9 40 E
Görlitz .... 51 10 N 15 00 E
Goslar .... 51 54 N 10 26 E
Gotha .... 50 57 N 10 43 E
Göttingen .... 51 32 N 9 56 E
Greifswald .... 54 06 N 13 23 E
Greiz .... 50 39 N 12 12 E
Grevesmühlen .... 53 52 N 11 11 E
Grossenhain .... 51 17 N 13 33 E
Guben (Wilhelm-Pieck Stadt Guben) .... 51 57 N 14 43 E
Güstrow .... 53 48 N 12 10 E
Gütersloh .... 51 54 N 8 23 E
Hagen .... 51 21 N 7 28 E
Hagenow .... 53 26 N 11 11 E
Halberstadt .... 51 54 N 11 03 E
Haldensleben .... 52 18 N 11 25 E
Halle .... 51 30 N 12 00 E
Halle-Neustadt .... 51 29 N 11 56 E
Hamburg .... 53 33 N 10 00 E
Hameln .... 52 06 N 9 21 E
Hamm .... 51 41 N 7 48 E
Hanau .... 50 08 N 8 55 E
Hannover .... 52 22 N 9 43 E
Harzgerode .... 51 38 N 11 09 E
Havelburg .... 52 49 N 12 05 E
Hechingen .... 48 21 N 8 59 E
Heidelberg .... 49 25 N 8 42 E
Heidenheim .... 48 41 N 10 09 E
Heilbronn .... 49 08 N 9 13 E
Helmstedt .... 52 14 N 11 00 E
Herford .... 52 08 N 8 41 E
Herne .... 51 33 N 7 13 E
Hildesheim .... 52 09 N 9 58 E
Hof .... 50 19 N 11 55 E
Homburg .... 49 19 N 7 20 E
Höxter .... 51 46 N 9 23 E
Hoyerswerda .... 51 26 N 14 15 E
Hürth .... 50 52 N 6 52 E
Idar-Oberstein .... 49 42 N 7 18 E
Ingolstadt .... 48 46 N 11 26 E
Iserlohn .... 51 22 N 7 42 E
Jena .... 50 56 N 11 35 E
Jülich .... 50 56 N 6 22 E
Kaiserslautern .... 49 27 N 7 45 E
Karl-Marx-Stadt, see Chemnitz
Karlsruhe .... 49 01 N 8 24 E
Kassel .... 51 19 N 9 30 E
Kaufbeuren .... 47 53 N 10 37 E
Kelheim .... 48 55 N 11 52 E
Kempten .... 47 43 N 10 19 E
Kerpen .... 50 52 N 6 41 E
Kesselsdorf .... 51 02 N 13 35 E
Kiel .... 54 20 N 10 08 E
Kleve .... 51 47 N 6 09 E
Koblenz .... 50 21 N 7 36 E
Köln, see Cologne
Königswinter .... 50 41 N 7 11 E
Konstanz .... 47 40 N 9 11 E
Korbach .... 51 17 N 8 52 E
Köthen .... 51 45 N 11 58 E
Krefeld .... 51 20 N 6 34 E
Kulmbach .... 50 06 N 11 27 E
Kunnersdorf .... 51 12 N 14 56 E
Landshut .... 48 32 N 12 09 E
Lauchhammer .... 51 30 N 13 48 E
Lehrte .... 52 23 N 9 58 E
Leinefelde .... 51 22 N 10 20 E
Leipzig .... 51 18 N 12 20 E
Lemgo .... 52 02 N 8 54 E
Leuna .... 51 19 N 12 01 E
Leverkusen .... 51 01 N 6 59 E
Limburg [an der Lahn] .... 50 23 N 8 03 E
Lindau .... 47 33 N 9 41 E
Lingen .... 52 31 N 7 19 E
Lippstadt .... 51 40 N 8 21 E
Löbau .... 51 06 N 14 40 E
Lörrach .... 47 37 N 7 40 E
Lorsch .... 49 39 N 8 34 E
Lübeck .... 53 52 N 10 42 E
Luckenwalde .... 52 05 N 13 10 E
Lüdenscheid .... 51 13 N 7 37 E
Ludwigsburg .... 48 54 N 9 11 E
Ludwigshafen .... 49 29 N 8 27 E
Lüneburg .... 53 15 N 10 24 E
Lünen .... 51 37 N 7 31 E
Lutter .... 51 59 N 10 16 E
Maasholm .... 54 41 N 9 59 E
Magdeburg .... 52 10 N 11 40 E
Mainz .... 50 00 N 8 15 E
Mannheim .... 49 29 N 8 28 E
Mansfeld .... 51 35 N 11 28 E
Marburg [an der Lahn] .... 50 49 N 8 46 E
Marl .... 51 39 N 7 05 E
Meerane .... 50 51 N 12 28 E
Meiningen .... 50 33 N 10 25 E
Meissen .... 51 09 N 13 29 E
Memmingen .... 47 59 N 10 10 E
Menden .... 51 26 N 7 48 E
Merseburg .... 51 22 N 12 00 E
Minden .... 52 17 N 8 55 E
Mittenwald .... 47 26 N 11 15 E
Moers .... 51 27 N 6 39 E
Mölln .... 53 38 N 10 41 E
Mönchengladbach .... 51 12 N 6 26 E
Mühlhausen .... 51 13 N 10 27 E
Mülheim [an der Ruhr] .... 51 26 N 6 53 E
Münden .... 51 25 N 9 41 E
Munich .... 48 09 N 11 35 E
Münster .... 51 58 N 7 38 E
Naumburg .... 51 09 N 11 49 E
Neu-Ulm .... 48 24 N 10 01 E
Neubrandenburg .... 53 34 N 13 16 E
Neumarkt [in der Oberpfalz] .... 49 17 N 11 28 E
Neumünster .... 54 04 N 9 59 E
Neunkirchen .... 49 21 N 7 11 E
Neuruppin .... 52 56 N 12 48 E
Neuss .... 51 12 N 6 42 E
Neustrelitz .... 53 22 N 13 05 E
Neuwied .... 50 26 N 7 28 E
Norden .... 53 36 N 7 12 E
Nordenham .... 53 30 N 8 29 E
Norderstedt .... 53 42 N 10 01 E
Nordhausen .... 51 31 N 10 48 E
Nordhorn .... 52 26 N 7 05 E
Nördlingen .... 48 51 N 10 30 E
Northeim .... 51 42 N 10 00 E
Nürnberg (Nuremberg) .... 49 27 N 11 05 E
Oberhausen .... 51 28 N 6 51 E
Oelsnitz .... 50 25 N 12 10 E
Offenbach .... 50 06 N 8 46 E
Offenburg .... 48 29 N 7 56 E
Oldenburg .... 54 18 N 10 53 E
Oldenburg .... 53 10 N 8 12 E
Oranienburg .... 52 45 N 13 14 E
Osnabrück .... 52 16 N 8 03 E
Osterholz-Scharmbeck .... 53 14 N 8 48 E
Osterode .... 51 44 N 10 11 E
Ottobrunn .... 48 04 N 11 41 E
Paderborn .... 51 43 N 8 46 E
Papenburg .... 53 04 N 7 24 E
Parchim .... 53 26 N 11 51 E
Passau .... 48 35 N 13 29 E
Peenemünde .... 54 08 N 13 47 E
Peine .... 52 19 N 10 14 E
Petersdorf .... 54 29 N 11 04 E
Pforzheim .... 48 53 N 8 42 E
Pirmasens .... 49 12 N 7 36 E
Pirna .... 50 58 N 13 56 E
Plauen .... 50 30 N 12 08 E
Potsdam .... 52 24 N 13 04 E
Prenzlau .... 53 19 N 13 52 E
Quedlinburg .... 51 47 N 11 09 E
Rathenow .... 52 36 N 12 20 E
Ravensburg .... 47 47 N 9 37 E
Recklinghausen .... 51 37 N 7 12 E
Regensburg .... 49 01 N 12 06 E
Reichenbach .... 50 37 N 12 18 E
Remagen .... 50 34 N 7 14 E
Remscheid .... 51 11 N 7 12 E
Rendsburg .... 54 18 N 9 40 E
Reutlingen .... 48 29 N 9 13 E
Rheine .... 52 17 N 7 27 E
Riesa .... 51 18 N 13 18 E
Rosenheim .... 47 51 N 12 08 E
Rostock .... 54 05 N 12 08 E
Rothenburg [ob der Tauber] .... 49 23 N 10 11 E
Rottweil .... 48 10 N 8 37 E
Rüdesheim .... 49 59 N 7 55 E
Rudolstadt .... 50 43 N 11 20 E
Rüsselsheim .... 50 00 N 8 25 E
Saalfeld .... 50 39 N 11 22 E
Saarbrücken .... 49 14 N 7 00 E
Saarlouis .... 49 19 N 6 45 E
Salzgitter .... 52 05 N 10 20 E
Salzwedel .... 52 51 N 11 09 E
Sangerhausen .... 51 28 N 11 18 E
Sankt Augustin .... 50 46 N 7 11 E
Sankt Ingbert .... 49 17 N 7 07 E
Schleswig .... 54 31 N 9 33 E
Schönebeck .... 52 01 N 11 45 E
Schönefeld .... 52 23 N 13 31 E
Schwäbisch Gmünd .... 48 48 N 9 47 E

Schwäbisch Hall . . 49 06 N 9 44 E
Schwedt . . . . . 53 04 N 14 18 E
Schweinfurt . . . . 50 03 N 10 14 E
Schwerin . . . . . 53 38 N 11 23 E
Seevetal . . . . . 53 24 N 9 58 E
Senftenberg . . . . 51 31 N 14 01 E
Siegen . . . . . 50 52 N 8 02 E
Sindelfingen . . . . 48 42 N 9 01 E
Singen . . . . . 47 46 N 8 50 E
Soest . . . . . 51 35 N 8 07 E
Solingen . . . . . 51 11 N 7 05 E
Sömmerda . . . . . 51 09 N 11 06 E
Sondershausen . . . 51 22 N 10 52 E
Sonneberg . . . . . 50 22 N 11 10 E
Speyer . . . . . 49 19 N 8 26 E
Spremberg . . . . . 51 33 N 14 22 E
Stade . . . . . 53 36 N 9 29 E
Stassfurt . . . . . 51 52 N 11 35 E
Stendal . . . . . 52 36 N 11 51 E
Stolberg . . . . . 50 46 N 6 14 E
Stralsund . . . . . 54 18 N 13 06 E
Straubing . . . . . 48 53 N 12 34 E
Strausberg . . . . . 52 35 N 13 53 E
Stuttgart . . . . . 48 46 N 9 11 E
Suhl . . . . . 50 36 N 10 42 E
Torgau . . . . . 51 34 N 13 00 E
Torgelow . . . . . 53 38 N 14 01 E
Triberg . . . . . 48 06 N 8 14 E
Trier . . . . . 49 45 N 6 38 E
Troisdorf . . . . . 50 49 N 7 10 E
Tübingen . . . . . 48 32 N 9 03 E
Tuttlingen . . . . . 47 59 N 8 49 E
Uelzen . . . . . 52 58 N 10 34 E
Ulm . . . . . 48 24 N 10 00 E
Vechta . . . . . 52 43 N 8 17 E
Verden . . . . . 52 55 N 9 14 E
Villingen-
Schwenningen . . . 48 04 N 8 27 E
Völklingen . . . . . 49 15 N 6 51 E
Walsrode . . . . . 52 54 N 9 40 E
Waren . . . . . 53 31 N 12 41 E
Warnemünde . . . . 54 10 N 12 05 E
Wedel . . . . . 53 35 N 9 43 E
Weiden [in der
Oberpfalz] . . . . 49 41 N 12 10 E
Weimar . . . . . 50 59 N 11 19 E
Weinheim . . . . . 49 33 N 8 40 E
Weissenfels . . . . 51 12 N 11 58 E
Weisswasser . . . . 51 31 N 14 38 E
Wernigerode . . . . 51 50 N 10 47 E
Wesel . . . . . 51 39 N 6 37 E
Wetzlar . . . . . 50 33 N 8 30 E
Weyhe . . . . . 52 58 N 8 48 E
Wiesbaden . . . . . 50 05 N 8 15 E
Wilhelm-Pieck-
Stadt Guben,
see Guben
Wilhelmshaven . . . 53 31 N 8 08 E
Wismar . . . . . 53 54 N 11 28 E
Wittenberg . . . . . 51 52 N 12 39 E
Wittenberge . . . . 53 00 N 11 45 E
Wolfen . . . . . 51 40 N 12 17 E
Wolfenbüttel . . . . 52 10 N 10 33 E
Wolfsburg . . . . . 52 26 N 10 48 E
Worms . . . . . 49 38 N 8 21 E
Wuppertal . . . . . 51 16 N 7 11 E
Würzburg . . . . . 49 48 N 9 56 E
Zehdenick . . . . . 52 59 N 13 20 E
Zeitz . . . . . 51 03 N 12 09 E
Zell . . . . . 50 02 N 7 11 E
Zerbst . . . . . 51 58 N 12 05 E
Zittau . . . . . 50 54 N 14 50 E
Zweibrücken . . . . 49 15 K 7 22 E
Zwickau . . . . . 50 44 N 12 30 E

Physical features and points of interest

Aller, *river* . . . . 52 57 N 9 11 E
Alps, *mountains* . . . 47 00 N 12 00 E
Alster, *river* . . . . . 53 34 N 10 01 E
Altmark, *region* . . . 52 45 N 11 30 E
Ammer, Lake . . . . 48 00 N 11 07 E
Baltic Sea . . . . . 54 30 N 13 00 E
Bavarian Alps,
mountains . . . . . 47 30 N 11 00 E
Bavarian Forest,
mountains . . . . . 49 00 N 12 40 E
Bavarian Forest
National
Park . . . . . 48 56 N 13 30 E
Berchtesgaden
National Park . . . . 47 32 N 12 57 E
Black Forest
(Schwarzwald),
mountains . . . . . 48 00 N 8 15 E
Bodensee, see
Lake Constance
Bohemian Forest,
mountains . . . . . 49 00 N 13 00 E
Brocken Peak . . . . 51 48 N 10 37 E
Chiem, Lake . . . . . 47 53 N 12 28 E
Constance, Lake
(Bodensee) . . . . 47 35 N 9 25 E
Danube (Donau),
river . . . . . 48 26 N 10 10 E
Dill, *river* . . . . . 50 33 N 8 29 E
Dollart Bay . . . . . 53 15 N 7 10 E
Donau,
see Danube
Dümmer, Lake . . . . 52 30 N 8 21 E
East Frisian
Islands . . . . . 53 44 N 7 25 E
Eider, *river* . . . . . 54 18 N 8 57 E
Eifel Upland . . . . . 50 15 N 6 40 E
Elbe, *river* . . . . . 53 50 N 9 00 E
Elbe Bridge . . . . . 53 33 N 9 55 E
Elbe-Havel
Canal . . . . . 52 24 N 12 23 E
Elde, *river* . . . . . 53 14 N 11 27 E
Elster, *river* . . . . . 50 17 N 12 14 E
Ems, *river* . . . . . 53 30 N 7 00 E
Emsland, *region* . . 52 50 N 7 20 E
Erzgebirge, see
Ore Mountains
Fehmarn, *island* . . . 54 28 N 11 08 E
Feld, Mount . . . . . 47 52 N 8 00 E
Fichtel Hills . . . . . 50 00 N 12 00 E
Fichtel Mountain . . . 50 26 N 12 57 E
Fläming Hills . . . . . 52 00 N 12 45 E
Franconian
Forest,
mountains . . . . . 50 18 N 11 36 E
Franconian Jura
(Fränkische Alb),
mountains . . . . . 49 10 N 11 30 E
Fulda, *river* . . . . . 51 25 N 9 39 E
Grosser Beer,
Mount . . . . . 50 39 N 10 45 E
Haardt
Mountains . . . . . 49 21 N 8 08 E
Harz Mountains . . . 51 45 N 10 30 E
Havel, *river* . . . . . 52 53 N 11 58 E
Helgoland,
island . . . . . 54 12 N 7 53 E
Helpter Hill . . . . . 53 29 N 13 36 E
Hohensaaten-
Friedrichsthal
Canal . . . . . 52 52 N 14 09 E
Hunsrück
Upland . . . . . 49 50 N 7 10 E
Iller, *river* . . . . . 48 23 N 9 58 E
Inn, *river* . . . . . 48 35 N 13 28 E
Isar, *river* . . . . . 48 48 N 12 59 E
Jade Bay . . . . . 53 27 N 8 12 E
Kaiserstuhl
Massif . . . . . 48 06 N 7 40 E
Kiel (Nord-
Ostsee)
Canal . . . . . 53 53 N 9 08 E
Königssee, *lake* . . . 47 33 N 12 58 E
Kummerow,
Lake . . . . . 53 49 N 12 52 E
Kyffhäuser
Mountains . . . . . 51 23 N 11 05 E
Lahn, *river* . . . . . 50 18 N 7 37 E
Lausitz (Lusatia),
region . . . . . 51 25 N 14 35 E
Lech, *river* . . . . . 48 44 N 10 56 E
Leine, *river* . . . . . 52 43 N 9 36 E
Lippe, *river* . . . . . 51 39 N 6 38 E
Lüneburg Heath,
region . . . . . 53 00 N 10 05 E
Main, *river* . . . . . 50 00 N 8 18 E
Main-Danube
(Main-Donau)
Canal . . . . . 49 54 N 10 54 E
Meckenburg,
physical region . . 53 30 N 12 00 E
Mecklenburg
Bay . . . . . 54 20 N 11 40 E
Mittelland Canal . . 52 16 N 11 41 E
Moselle (Mosel),
river . . . . . 50 22 N 7 36 E
Mulde, *river* . . . . . 51 52 N 12 15 E
Müritz, Lake . . . . . 53 25 N 12 42 E
Münsterland,
region . . . . . 51 50 N 8 00 E
Naab, *river* . . . . . 49 01 N 12 02 E
Nahe, *river* . . . . . 49 58 N 7 57 E
Neckar, *river* . . . . 49 31 N 8 26 E
Neisse, *river* . . . . 52 04 N 14 46 E
Niedersachsisches
Wattenmeer
National Park . . . 53 45 N 7 30 E
Nord-Ostsee
Canal, see
Kiel Canal
Norderney,
island . . . . . 53 43 N 7 15 E
North Frisian
Islands . . . . . 54 50 N 8 12 E
North Sea . . . . . 54 30 N 6 30 E
Oberlausitz,
region . . . . . 51 10 N 14 20 E
Oden Forest,
region . . . . . 49 38 N 8 50 E
Oder, *river* . . . . . 53 32 N 14 38 E
Oder-Spree
Canal . . . . . 52 23 N 13 41 E
Ore Mountains
(Erzgebirge) . . . . 50 30 N 13 10 E
Ortenau, *area* . . . . 48 27 N 8 00 E
Oste, *river* . . . . . 53 51 N 8 59 E
Peene, *river* . . . . . 54 09 N 13 46 E
Plau, Lake . . . . . 53 28 N 12 19 E
Prignitz, *region* . . . 53 00 N 12 00 E
Recknitz, *river* . . . . 54 15 N 12 28 E
Rechnitz, *river* . . . . 49 28 N 10 59 E
Regen, *river* . . . . . 49 02 N 12 06 E
Reichenau,
island . . . . . 47 41 N 9 04 E
Rhine, *river* . . . . . 51 52 N 6 02 E
Rhön Mountains . . . 50 30 N 10 00 E
Rothaar Hills . . . . . 51 05 N 8 15 E
Rügen, *island* . . . . 54 25 N 13 24 E
Ruhr, *river* . . . . . 51 27 N 6 44 E
Saale, *river* . . . . . 51 57 N 11 55 E
Saar, *river* . . . . . 49 11 N 7 06 E
Sauerland,
region . . . . . 51 20 N 7 50 E
Schleswig-
Holstein
National Park . . . 54 19 N 8 57 E
Schnee, Mount . . . . 50 03 N 11 51 E
Schwäbische Alb,
see Swabian Jura
Schwarze Elster,
river . . . . . 51 49 N 12 51 E
Schwarzwald,
see Black Forest
Schwerin, Lake . . . . 53 43 N 11 28 E
Seven Hills . . . . . 50 40 N 7 14 E
Spessart, *region* . . . 49 55 N 9 25 E
Spree, *river* . . . . . 52 32 N 13 13 E
Starnberg, Lake . . . 47 55 N 11 19 E
Steinhude, Lake . . . 52 28 N 9 20 E
Swabian Jura
(Schwäbische
Alb), *mountains* . . 48 25 N 9 30 E
Sylt, *island* . . . . . 54 54 N 8 20 E
Taunus
Mountains . . . . . 50 15 N 8 15 E
Tegern, Lake . . . . . 47 43 N 11 44 E
Teutoburg
Forest,
mountains . . . . . 52 05 N 8 20 E
Thuringia
(Thüringen),
region . . . . . 51 00 N 11 00 E
Thuringian
Forest,
mountains . . . . . 50 40 N 10 50 E
Uckermark,
region . . . . . 53 07 N 13 30 E
Vogels Mountain . . . 50 31 N 9 14 E
Vogtland,
region . . . . . 50 30 N 12 15 E
Warnow, *river* . . . . 54 06 N 12 09 E
Wasser
Mountain . . . . . 50 30 N 9 56 E
Werra, *river* . . . . . 51 26 N 9 39 E
Weser, *river* . . . . . 53 32 N 8 34 E
Wester Forest,
mountains . . . . . 50 40 N 7 55 E
Wörnitz, *river* . . . . 48 42 N 10 45 E
Zugspitze, *peak* . . . 47 25 N 10 59 E

ments, before rising to the smaller Oden Forest. For the most part, however, southern Germany consists of scarplands, mainly of Triassic age (208 to 245 million years old). The work of erosion on eastward-dipping strata has left the sandstones standing out as west- or northwest-facing scarps, overlooking valleys or low plateaus of clays or Muschelkalk (Triassic limestone formed from shells). The sequence of Triassic rocks ends south and east against the great Jurassic scarp of the Swabian Alp (Schwäbische Alb), rising to more than 3,300 feet (1,000 metres), and its continuation, the lower Franconian Alp (Fränkische Alb). Large parts of the plateaus and lowlands in the eastern region are covered with loess, but the massive Bunter Sandstone fringing the Black Forest and the Keuper scarp are mainly wooded. West of the Rhine there are again wide stretches of forested Bunter Sandstone, with more open country in the Saar region and along the foot of the Hunsrück Upland.

The barrier arc. The open land of southern Germany ends against a great barrier arc of Hercynian massifs and forested sandstone plateaus. In the west the Rhenish Uplands (Rheinisches Schiefergebirge) consist mainly of resistant slates and shales. The complex block is tilted generally northwestward, with a steep fault-line scarp in the south. The intensely folded rocks are planed off by erosion surfaces that give the massif a rather monotonous appearance, broken only by quartzite ridges, especially in the south, where the Hunsrück rises to 2,684 feet (818 metres) and the Taunus to 2,887 feet (879 metres).

The valleys

The valleys are a different world. They range from narrow forested slots—a great hindrance to passage—to the spectacular gorge of the Rhine, the most important natural routeway through the barrier arc. The most dramatic section of the gorge runs from Bingen to the vicinity of Koblenz; hilltop castles look down over vineyards to picturesque valley towns. In this section is the Lorelei rock, from which a legendary siren is supposed to have lured fishermen to their death on the rocks.

Until highways were constructed over the plateau tops, access to the uplands was difficult. The landscape gained

some variety from past volcanic activity responsible for the eroded volcanic necks of the Seven Hills (Siebengebirge), across from Bonn, the flooded craters and cinder cones of the Eifel Upland, and the sombre basalt flows of the Westerwald. Westward the Rhenish Uplands continue into Belgium as the Ardennes. In the Carboniferous Period (from 360 to 286 million years ago) when the Hercynian Uplands were still young folded mountains, great deltaic swamps developed to the north and south; these were the basis of the great Ruhr coalfield and the smaller Aachen and Saar fields.

The Bohemian Massif

The eastern end of the barrier arc is buttressed by the great and complex Bohemian Massif, which Germany shares only marginally. On the southwestern fringe of the massif, German territory includes the remote and thinly populated Bohemian Forest and the Bavarian Forest. Across from Czechoslovakia are the Ore Mountains (Erzgebirge), where the centuries-old mining tradition was still continued in the period of the German Democratic Republic. The Bohemian Massif is prolonged northwestward by the long spur of the Thuringian Forest (Thüringer Wald), which separates the scarplands of northern Bavaria from the Thuringian Lowland. The barrier arc is completed by the great eroded cone of the Vogelberg (2,536 feet [773 metres]), the volcanic Rhön, and the forested Bunter Sandstone plateaus of northern Hesse. The Rhine Rift Valley continues northward through Hesse, with a series of discontinuous basins filled with Tertiary sediments that allow a slightly difficult traverse to the North German Plain.

The northern fringe of the Central Uplands. North of the upland barrier there are a number of regions, generally of folded limestones, sandstones, and clays, that mark the transition to the expanse of the North German Plain. Balanced on either side of the plateau of Hesse are two basins of subdued scarpland relief, the Westphalian Basin to the northwest and the Thuringian Basin to the southeast, both of them partially invaded by glacial outwash from the North German Plain. Hesse and the Westphalian Basin are succeeded northward by the hills of Lower Saxony. The breakthrough of the Weser River into the North German Plain at the Porta Westfalica, south of Minden, is overlooked by the giant monument of the emperor William I (built in 1896). North of the Thuringian Basin is one of the smaller Hercynian massifs, the Harz, which reaches an elevation of 3,747 feet (1,142 metres) in the Brocken Peak.

The North German Plain, or Lowland. Less than 90 miles broad in the west, the North German Plain widens eastward across the whole of northern Germany. Although relief is everywhere subdued, there is considerable variety of landscape in terms of detail and much natural beauty. Unconsolidated Tertiary deposits, gravels, sands, and clays, with overlying glacial drift, have buried the previous landscape of secondary rocks. These make only two brief appearances, in the chalk cliffs of the island of Rügen in the Baltic Sea and in the cliffs of Triassic Bunter Sandstone of the island of Helgoland, located some 40 miles northwest of Cuxhaven in the North Sea. In Tertiary times large swamps developed, and the underlying deposits of lignite (brown coal) are presently mined in Saxony, in Lower Lusatia (Niederlausitz), and west of the city of Cologne.

The North German Plain is divided into contrasting eastern and western portions, the division marked approximately by the Elbe valley. The northern and eastern regions were molded by southward-moving ice sheets in the last (Weichsel, or Vistula) glaciation. The advancing ice sheets pushed up material that remains today as terminal moraines, stretching across the country in a generally southeast to northwest direction and rising to some 500 feet above the general level. Within the terminal moraines the decay of the ice sheets typically left behind sheets of till (ground moraine), which have mainly been cleared of their heavy soils for agriculture. They are studded with ponds, often resulting from the decay of buried "dead ice," and littered with boulders of all sizes brought by the ice from Scandinavia. In a region lacking in stone these boulders were used as building material and are to be found forming the walls of the oldest churches. Outside the moraines, meltwater laid down sheets of outwash sands, which, offering poorer soils, are frequently forested. A feature of the present-day moraine country is the existence of large, long, and branching lake systems, usually believed to have been formed by water moving under the ice sheets.

The terminal moraines

The unique character of the region east of the Elbe is further enhanced by the fact that the ice sheets of the last glaciation coming from the north blocked the river's natural flow to the Baltic, forcing it to escape laterally around the margin of the ice toward the North Sea; the river cut a deep trench as it did so. The landscape in the western portion of the plain tends to be monotonous. Much of it was formerly heath; the few patches that have escaped afforestation, agricultural improvements, or damage caused by military training have a wistful beauty, especially when the heather is in bloom. Wilseder Berg (554 feet [169 metres]), a fragment of a former moraine, is the highest elevation in the Lüneburg Heath, a plateau extending on

E. Streichan—Shostal Associates/Superstock



Barge on the Rhine River, with vineyards in the background, at the town of Kaub, Rhineland-Westphalia, Ger.

a morainic belt between Hamburg and Hannover. Toward the maritime northwest, large areas of peat bogs have been reclaimed for agriculture. The southern edge of the plain extending to the Thuringian Basin is marked by a belt of mainly windblown loess.

The coasts. The western and eastern coastlines vary considerably in their forms. The coast of the North Sea continues the type familiar in the northern Netherlands; an offshore bar, crowned with sand dunes, has been shattered and left as the chain of the East Frisian Islands off the coast of Lower Saxony and the North Frisian Islands off the Schleswig-Holstein portion of the Jutland Peninsula. These islands form a favourite vacation area in summer, even if it is often necessary to seek shelter from the wind in high-backed wicker beach chairs. The sea has encroached upon the land behind the islands, forming tidal flats (known as *Wattenmeer*),which become exposed at low tide. The coast is broken by the estuaries of the Elbe, Weser, and Ems rivers and by drowned inlets such as the Jade and Dollart bays.

The *Wattenmeer*

Along the Baltic coast, the boulder-clay plains shelve rather tamely beneath the sea. However, the typically varied relief of minor moraines, depressions, and other glacial features gives suffcient diversity to the coastline. In Schleswig-Holstein long inlets (fjords), carved by water moving beneath the ice sheets, extend to the sea. Farther east the coast gains in complexity; there are peninsulas and sea inlets known as *Bodden,* and sandy beach bars dominate the landscape. Several islands line the shore, including Usedom, Hiddensee, Poel, and Rügen, Germany's largest island.

The Alps and the Alpine Foreland. Very small portions of the outer limestone (or calcareous) Alps extend from Austria into Germany. From west to east these are the Allgäuer Alps, the Wetterstein Alps—with Germany's highest mountain, the Zugspitze (9,718 feet)—and the Berchtesgadener Alps. Like the North German Plain, the Alpine Foreland is fundamentally a depression filled with Tertiary gravels, sands, and clays, which are derived from the Alpine orogeny. But, in contrast to the North German Plain, the Tertiary deposits are more visible on the surface. Along the foot of the limestone Alps, but particularly in the Allgäuer Alps in the west, the older Tertiary deposits (flysch, molasse) were caught up in the later stages of the Alpine folding, forming a pre-Alpine belt of hills and low mountains consisting mainly of sandstone. The Tertiary sands and clays also emerge at a much lower elevation in the northeast, forming a subdued landscape.

Glaciers emerging from the main Alpine valleys formed lobes stretching some 20 to 35 miles into the plain. Crescentic moraines mark the points where the lobes came to rest; within the moraines are irregular deposits of till and many lakes. Outside the moraines, floodwaters deposited sheets of outwash gravel, which extend as river terraces along the courses of tributaries flowing north to the Danube. The Alps and the Bavarian lakes are among Germany's most favoured tourist areas.

DRAINAGE AND SOILS

Drainage. Most German rivers follow the general north-to-northwest inclination of the land, eventually entering the North Sea. The major exception to the rule is the Danube; it rises in the Black Forest and flows eastward, marking approximately the boundary between the Central German Uplands and the Alpine Foreland. It draws upon a series of right-bank Alpine tributaries, which, through reliance on spring and summer snowmelt, make its regime notably uneven. Further exceptions are the Altmühl and the Naab, which follow a southerly direction until becoming north-bank tributaries of the Danube, and the Havel, which flows south, west, and north before emptying into the Elbe River. River flow relates mainly to climate, albeit not in a simple way; for example, in all but Alpine Germany, maximum river flow occurs in winter when evaporation is low, although in the lowlands the peak rainfall is in summer.

The most majestic of the rivers flowing through Germany is the Rhine. It has its source in east-central Switzerland, flows west through Lake Constance (Bodensee), skirting the Black Forest to turn northward to flow across the Central German Uplands. Below Bonn the Rhine emerges into a broad plain, and west of Emmerich it enters The Netherlands to issue into the North Sea. It is of great advantage that the Rhine belongs to two types of river regime. Rising in the Alps, the Rhine profits first from the extremely torrential Alpine regime, which causes streams to be swollen by snowmelt in late spring and summer. Then, by means of its tributaries, the Neckar, Main, and Moselle (German: Mosel), the Rhine receives the drainage of the Central German Uplands and the eastern part of France, which contributes to a maximum flow during the winter. As a result, the river has a remarkably powerful and even flow, a physical endowment that caused it to become the busiest waterway in Europe. Only in occasional dry autumns are barges unable to load to full capacity to pass the Rhine gorge.

The Rhine

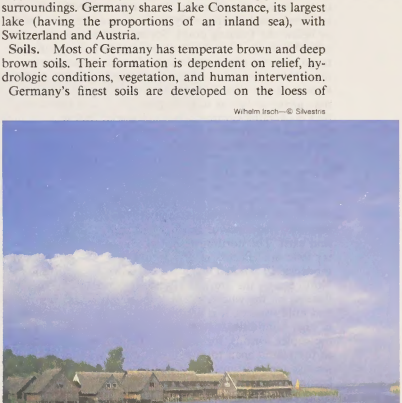
The Weser and Elbe rise in the Central German Uplands, crossing the North German Plain to enter the North Sea. The northward-flowing Oder (with its tributary, the Neisse) passes through the northeastern part of the country and a small section of Poland before emptying into the Baltic Sea. The navigation of these rivers is often adversely affected by low water in the summer and by ice in the winter, which increases eastward.

River courses in the northern lowlands have a notably trellised pattern—rivers follow the ice-margin stream trenches (*Urstromtäler*) carved outside the fringes of the retreating ice sheets before breaking through the next moraine ridge to the north. This pattern greatly facilitated the cutting of canals linking the Rhine River with Berlin and the Elbe and Oder rivers.

Germany has relatively few lakes. The greatest concentration comprises the shallow lakes of the postglacial lowland of the northeast. The largest natural lake of the region is Lake Müritz (44 square miles [114 square kilometres]) in the Weichsel glacial drift of Mecklenburg-West Pomerania. In addition to Lakes Dümmer and Steinhude in Lower Saxony, a few small lakes of glacial origin dot Schleswig-Holstein. The remainder of Germany's lakes are concentrated at the extreme southeastern corner of Upper Bavaria, many of these in outstandingly beautiful surroundings. Germany shares Lake Constance, its largest lake (having the proportions of an inland sea), with Switzerland and Austria.

Soils. Most of Germany has temperate brown and deep brown soils. Their formation is dependent on relief, hydrologic conditions, vegetation, and human intervention. Germany's finest soils are developed on the loess of

Wilhelm Irsch—© Silvestris



Fishermen's huts (*Fischerhäuser*) on the west coast of Lake Müritz near Röbel, Mecklenburg-West Pomerania.

the northern flank of the Central German Uplands, the Magdeburg Plain, the Thuringian Basin and adjoining areas, the Rhine valley, and the Alpine Foreland. They range from black to extremely fertile brown soil types, and most of them are arable land under cultivation. The till (ground moraine) of the North German Plain and Alpine Foreland has heavy but fertile soil. Brown soil covers much of the Central German Uplands and is used for agriculture and grazing. With increasing elevation, soils are suitable only for grazing or forestation. In the northern plains the soil types are sand, loam, and brown podzols, which are heavily leached of mineral matter and humus by deforestation and grazing. Along the North Sea littoral in the northwest there are some extensive areas of sand, marsh, and mud flats that are covered with rich soil suitable for grazing and growing crops.

The remainder of German soil types, because of the preponderance of mountainous and forested areas, range from sand to loam, from loam to clay, and from clay to rocky outcrops. Timber production thrives where the land is all but unarable, and viticulture in the southern regions flourishes in an otherwise inhospitable type of soil.

(T.H.El.)

CLIMATE

Germany is favoured with a generally temperate climate, especially in view of its northerly latitudes and the distance of the larger portions of its territory from the warming influence of the North Atlantic Current. Extremely high temperatures in the summer and deep, prolonged frost in the winter are rare. These conditions, together with a more than abundant and well-distributed amount of rainfall, afford ideal conditions for raising crops. As throughout western Europe in general, however, Germany's climate is subject to quick variations when the warm westerly winds from the Atlantic Ocean collide with the cold air masses moving in from northeastern Europe. Whereas in the open coastlands near the North and Baltic seas the maritime component prevails, continental elements gain in importance moving toward the east and southeast.

Variations in seasons

The seasons, year by year, are subject to great variations: winters may be unusually cold or prolonged, particularly in the higher elevations in the south, or mild, with the temperatures hovering only two or three degrees above or below the freezing point. Spring may arrive early and extend through a hot, rainless summer to a warm, dry autumn with the threat of drought. In other years, spring—invariably interrupted by a frosty lapse in May, popularly known as *die drei Eisheiligen* ("the three ice saints")—may arrive so late as to be imperceptible and be followed by a cool, rainy summer. One less agreeable feature of the German climate is an almost permanent overcast in the cool seasons, only infrequently accompanied by precipitation; it sets in toward the latter part of autumn and lifts as late as March or April. Thus, for months on end, virtually no sunshine may appear.

Despite the generally temperate climate for the country as a whole, specific regional patterns associated with temperature, frequency of sunshine, humidity, and precipitation exist. The northwestern and lowland portions of the republic are affected chiefly by the uniformly moist air, moderate in temperature, that is carried inland from the North Sea by the prevailing westerly winds. While this influence, on the whole, affords moderately warm summers and mild winters, it is accompanied by the disadvantages of high humidities, extended stretches of rainfall, and, in the cooler seasons, fog. Precipitation diminishes eastward as the plains open toward the Eurasian interior, and the average temperatures for the warmest and coldest months become more extreme. The hilly areas of the central and southwestern regions and, to an even greater degree, the upland and plateau areas of the southeast are subject to the more pronounced ranges of hot and cold from the countervailing continental climate. The mountains have

Mountain climate

a wetter and cooler climate, with westward-facing slopes receiving the highest rainfall from maritime air masses. At a station on the Brocken in the Harz Mountains, annual precipitation reaches 63 inches (160 centimetres) at an altitude of 3,747 feet. The sheltered lee slopes and basins have, by contrast, rainfall that is extremely low (Aisleben receives 17 inches annually) and hot summers (July mean temperatures above 64° F [18° C]), necessitating crop irrigation. Southeastern Germany may intermittently be the coldest area of the country in the winter; but the valleys of the Rhine, Main, Neckar, and Moselle rivers may also be the hottest in the summer. Winters in the North German Plain tend to be consistently colder, if only by a few degrees, than in the south, largely because of winds from Scandinavia. There is also a general decrease of winter temperature from west to east, with Berlin having an average temperature in January of 31.5° F (−0.3° C).

One anomaly of the climate of Upper Bavaria is the occasional appearance of warm, dry air passing over the northern Alps to the Bavarian Plateau. These mild winds, known as foehns (*Föhn*), can create an optical phenomenon that makes the Alps visible from points where they normally would be out of sight, and they also are responsible for the abrupt melting of the snow.

The foehns

Annual mean precipitation varies according to region. It is lowest in the North German Plain, where it ranges from 20 to 29 inches; in the Central German Uplands it ranges from 28 to 59 inches; and in the Alpine regions, up to and exceeding 78 inches.

PLANT AND ANIMAL LIFE

Since Germany is a somewhat arbitrary south-north slice across central Europe, it does not have vegetation and animal life greatly different from that of neighbouring countries. Before being settled, Germany was almost totally forested, except for a few areas of marsh. At present it would be hard to find any truly natural vegetation; not only the cultivated areas but the forests, which cover 20 percent of the land area, are man-made.

Plants. After the Ice Age the loess areas were covered by oak and hornbeam forests, which, however, have largely disappeared. The sand spreads of the North German Plain were originally covered by a predominantly mixed oak-birch woodland. They were cleared and replaced by heather (*Calluna vulgaris*) for sheep grazing, with associated soil erosion. In the 19th century artificial fertilizer was introduced to improve some of this land for agriculture, and large stretches were forested, mainly with Scotch pine (*Pinus sylvestris*). The Central Uplands are traditionally the domain of the beech (*Fagus sylvatica*), a tree with a leaf canopy so dense that few plants can survive beneath it. Although beech trees survive well on the poor soils covering the limestones and the Bunter Sandstone, they have been increasingly replaced by pine in the lowlands and spruce in the uplands. Other conifers, such as the Douglas and Sitka spruce, Weymouth pine, and Japanese larch, were introduced in recent years. In the highest elevations of the Alps, mixed forests and pasture provide grazing for cattle. German forests have suffered greatly from acid rain pollution, generally blamed on emissions (of sulfur dioxide and nitrogen oxide) from power plants and industrial operations. Damage has also been severe in southeastern Germany, near the Ore Mountains, which border on the Czechoslovakian lignite-burning industries.

Animals. The vast tracts of forest and mountainous terrain, with only scattered habitation, contribute to a surprising variety of wildlife in so densely populated and highly developed a country. Game animals abound in most regions—several varieties of deer, quail, and pheasant and, in the Alpine regions, the chamois and ibex—and their numbers are protected by stringent game laws. The wild boar population, which soared after World War II because of restrictions on hunting, has now been reduced so that it no longer represents a danger to people and crops. The hare, a favoured game animal, is ubiquitous. Although in the wild the bear and the wolf are now extinct in the republic, the wildcat has had a resurgence in the postwar years, especially in the Eifel and Hunsrück regions and in the Harz Mountains. The lynx has reappeared in the areas near the Czechoslovakian border, and the elk and wolf are occasional intruders from the east. The polecat, marten, weasel, beaver, and badger are found in the central and southern uplands, and the otter and wildcat are among the rarer animals of the Elbe basin. Common reptiles include

salamanders, slowworms, and various lizards and snakes, of which only the adder is poisonous.

Germany has several internationally recognized bird reserves. The tidal flats of Lower Saxony (Niedersächsisches Wattenmeer) and the Schleswig-Holstein national parks along the North Sea coast, the lakes of the Mecklenburg plains, and glacially formed lakes of the North German Plain are vital areas for the European migration of ducks, geese, and waders. The Lüneburg Heath, an excellent nature reserve, is a haven for various species of plants, birds, insects, and reptiles. The rare white-tailed eagle can be found in the lakes of the North German Plain, whereas the golden eagle can be seen in the Alps. White storks have decreased in number, but they can still be seen, perched on enormous piles of sticks on chimneys or church towers in areas where unpolluted and undrained marsh is still to be found.

SETTLEMENT PATTERNS

Rural settlement. The most striking feature of the rural settlement pattern in western Germany is probably the concentration of farmyards into extremely large villages, known as *Haufendörfer.* These villages are surrounded by unenclosed fields divided into often hundreds of strip-like units. The *Haufendorf* is particularly characteristic of Hesse and southwestern Germany, areas that have a tradition of partible inheritance. During periods of population pressure, land holdings—as well as farmhouses and farmyards—were repeatedly divided on inheritance, becoming smaller and more fragmented all the time. As a result, villages became increasingly huddled and chaotic. In areas with a tradition of undivided inheritance, such as Bavaria and Lower Saxony, the holdings—the individual field parcels and the farmhouses—remained larger.

Village forms

During the period of high medieval prosperity in the 11th to 13th centuries, population pressure brought about two further developments: the advance of peasant settlements into the forests, where isolated farms and hamlets were the usual settlement form, and the colonization of the predominantly Slav lands beyond the line of the Elbe and Saale rivers. Since this was an organized colonization, under the control of the lords and their agents, the haphazard structure of the western *Haufendorf* could be streamlined into a few well-planned forms. Thus farmhouses in the eastern regions were customarily arranged along either a single village street (*Strassendorf*) or an elongated green, on which stood the church (*Angerdorf*); long unfenced strips of land were allotted at right angles to the road or green.

The evolution of rural settlement has not been uniform; there have been phases of advance and retreat. In particular, the decline of medieval prosperity, accompanied by the Black Death, led to a stage of retreat, in which many hundreds of villages—the so-called "lost villages"—were abandoned in western Germany. In eastern Germany, lords often appropriated deserted farms and added them to their own land, thus initiating that characteristic feature of the area beyond the Elbe, the large Junker estate farm (*Gut*).

After World War II the organization of agriculture and settlement in the two Germanys diverged considerably. The western direction was one of evolution, with federal and state governments giving large subsidies to improve the existing structure. Land consolidation was one method: scattered strips were regrouped to form larger holdings, and in some places farmers were moved to new farmsteads dispersed outside the villages. An increased average size of holding was associated with a massive movement of people out of agriculture. But instead of leaving for the cities, as happened in the 19th century, people mostly remained in their existing homes and commuted to work; the manure tank in front of the house became a rose bed, and the barn a garage. Part-time "Sunday" farming remained, but on a reduced scale. Land was actually left uncultivated by the new urban workers (social fallow); some of it was taken up by the few remaining full-time farmers, but marginal areas were abandoned, being either afforested or reverting to rough grass and scrub.

Developments after World War II

In the former East Germany, the Junker estates were confiscated and either divided among peasants or turned into state farms. This development was only the first stage in a process of collectivization; from 1958 to 1960 private holdings were regrouped under heavy political pressure into vast "cooperative" farms. New buildings marked the introduction of mechanized cultivation or large-scale animal husbandry, and multistory apartments and community centres reflected a politically inspired attempt to create a new concept of rural life. Now all is once more in transformation.

Urban settlement. From medieval times onward Germany was politically fragmented, with numerous states competing with one another to develop lucrative market

Otto Stadler—© Silvestris



Village built along a single street (*Strassendorf*): Stolberg, Saxony-Anhalt.

centres and to create capitals, large and small; as a result, the country inherited a profusion of towns and cities. Most of these remained frozen within their circuit of walls until the 19th century; only the larger princely capitals, such as Berlin or Munich, developed distinctive government quarters in the early modern period. The great urban explosion came late in the 19th century. Because industrialization was largely linked to the development of the railways, urban expansion was not confined to areas near the coalfields, such as the Ruhr region, but was distributed among many cities. Typically the new urban workers were herded into dismal five-story apartment blocks built on a monotonous grid of straight-line streets. Today, especially in eastern Berlin and cities such as Halle and Leipzig, these blocks present an urgent problem of urban renewal.

World War II was followed by a period of rapid urban growth as evacuees returned to the bombed cities. After 1949 contrasting government policies of the two Germanys, however, led to divergent development. In the West many people abandoned the old city cores in favour of widespread suburbs and urbanized villages within commuting range. Thus, in many agglomerations, notably in the Ruhr region, population loss has been associated with peripheral gain. By contrast, the East German government pursued a policy of population concentration, whereby people were moved into concentrated peripheral settlements of 50,000 to 100,000, consisting of uniform prefabricated high-rise housing blocks.

The people

The German-speaking peoples—to whom must be accounted not only the inhabitants of Germany but also those of Austria, Liechtenstein, the major parts of Switzerland and Luxembourg, small portions of France and Italy, and the remnants of German communities in eastern Europe—are extremely heterogeneous in their ethnic origins, in their dialectal divisions, and in their political and cultural heritage, in which the split between Protestant and Roman Catholic has played a significant role since the Protestant Reformation and Catholic Counter-Reformation in the 16th century.

A characteristic of Germany, throughout its history, has been the lack of clearly defined geographic boundaries, particularly on the great lowland of northern Europe; both the area occupied by the German peoples and the boundaries of the German state (at such times as it existed) have fluctuated constantly. The German people appear to have originated on the coastal region of the Baltic Sea and in the Baltic islands in the Bronze and early Iron ages. From about 500 BC they began to move southward, crushing and absorbing the existing Celtic kingdoms; from 58 BC onward they clashed along the line of the Rhine and Danube rivers with the power of Rome. With the fall of the Roman Empire, German peoples, predominantly under Frankish tribal leadership, closely settled a large area west of the Rhine River in what is still German territory; they also penetrated deeply into Belgium and areas that later became France. The Merovingian and Carolingian empires knew no distinction between what are now France and western Germany; it is understandable that Charlemagne (Karl der Grosse) is recognized as an important figure in the history of both countries.

Origins of the German people

The weakness of Charlemagne's successors was revealed in their inability to deal with the waves of savage invaders that poured into the empire at the end of the 9th century. In despair, people turned to local leaders able to offer protection. In the German heartland, the old tribal divisions still retained their validity, and the tribes looked for defense to an army of their own people, led by a duke. (The names of these dukedoms are still in use for some of the states [*Länder*] of Germany today, notably Bavaria, Thuringia, and [Lower] Saxony.) In the 10th and 11th centuries they were brought under the power of a single monarch, but this precocious centralization did not survive. The dukedoms were progressively subdivided until Germany became notorious for its *Kleinstaaterei*—its swarm of frequently tiny states, each with its court borne on the backs of the peasantry. The states, and particularly their boundaries, were of considerable social and economic significance, introducing contrasts that are in part still perceptible today.

The rise of France as a centralized power extinguished most of Germanic control west of the Rhine, a process facilitated by German divisions; dialects of German remain in use in France only in Alsace and parts of Lorraine. However, driven by population pressure during the Middle Ages, Germans not only cleared large areas of forest for the expansion of cultivation but extended their settled area far to the east. From about 800 in the south, and about two centuries later in the centre and north, the Germans moved east in an advance that divided into three prongs: down the Danube through Austria, north of the Central German Uplands through Silesia, and along the Baltic shore. Between the prongs were the partially isolated Slav areas of Bohemia and Poland; this development held the potential for conflict that lasted until the 20th century. Islands of German people were at various times established beyond the continuously settled area as far as the Volga. The tenacity of these groups in retaining their German language and culture through the centuries was remarkable.

Eastward expansion

The German Empire created in 1871 had not included all German-speaking peoples; in particular, the Germans of the Austro-Hungarian Empire were excluded from the new Reich, while Switzerland, with its majority of German speakers, has maintained its independence to this day. After World War I large numbers of Germans who had lived under German or Austrian rule found themselves either in France—Alsace and Lorraine, German since 1871, having been returned to French control—or in the states created by the Treaty of Versailles in 1919, notably Poland and Czechoslovakia. The presence of German ethnic minorities in these countries was later used by Adolf Hitler as an excuse for military occupation. After World War II the German populations were largely expelled from Czechoslovakia and Poland; in this way the distribution of German-speaking people came more nearly to coincide with the boundaries of the German state, although Austria and German-speaking Switzerland still remained outside.

ETHNIC STRUCTURES

The Germans, in their various changes of territory, inevitably intermingled with other peoples. In the south and west they overran Celtic peoples, and there must at least have been sufficient communication for them to adopt the names of such continuing features as rivers and hills: the names Rhine, Danube, and Neckar, for example, are thought to be of Celtic origin. Similarly, in occupying the Slav lands to the east, they seem to have taken over and reorganized the Slav people along with their established framework of rural and urban settlements, many of which, including numerous physical features, still bear names of Slavic origin. The same is true of family names. In addition, large numbers of immigrants have added to the mixture: French Huguenots at the end of the 16th century, Polish mineworkers of the Ruhr at the end of the 19th, White Russians in Berlin after the revolution of 1917, and stateless "displaced persons" left behind by World War II.

Prior to the 1950s there were few ethnic minorities in Germany. A population of Slavic-speaking Sorbs (Wends), variously estimated at between 30,000 and an improbable 100,000, have survived in the Lusatia (Lausitz) area, between Dresden and Cottbus, while a small number of Danish speakers are still to be found in Schleswig-Holstein, even after the Versailles boundary changes there. Of the "guest workers" (*Gastarbeiter*) and their families, introduced from the mid-1950s onward, the largest group are the Turks; they are distinct in culture and religion, but they do not occupy a contiguous region, being spread throughout German cities. Even more culturally distinct groups have been added by asylum seekers from countries such as Sri Lanka and Vietnam, and the opening of the eastern frontiers has brought many more (see below).

Ethnic minorities

LINGUISTIC DIVERSITY

The dialectal divisions of Germany, once of conspicuous significance for the ethnic and cultural distinctions they

implied, persist despite leveling and standardizing influences such as mass education and communication and despite internal migration and the trend among the younger, better-educated, and more-mobile ranks of society to speak a standard, "accentless" German. The repository of differences in dialect now lies more with the rural populace and the longtime native inhabitants of the cities.

Standard German itself is something of a hybrid language in origin, drawn from elements of the dialects spoken in the central and southern districts but with the phonetic characteristics of the north predominating. Indeed, the pronunciation of standard German is an arbitrary compromise that gained universal currency only in the late 19th century. Even today the most "accent-conscious" of the well-educated speak with the coloration of their native district's dialect, especially if they are from the southern regions.

The dialects of German

The three major dialectal divisions of Germany coincide almost identically with the major topographic regions: the North German Plain (Low German), the Central German Uplands (Central German), and the southern Jura, Danube basin, and Alpine districts (Upper German). Of the Upper German dialects, the Alemannic branch in the southwest is subdivided into Swabian, Low Alemannic, and High Alemannic. Swabian, the most widespread and still-ascending form, is spoken to the west and south of Stuttgart and as far east as Augsburg. Low Alemannic is spoken in Baden-Württemberg and Alsace, and High Alemannic is the dialect of German-speaking Switzerland. The Bavarian dialect, with its many local variations, is spoken in the areas south of the Danube River and east of the Lech River and throughout all of Austria, except in the state of Vorarlberg, which is Swabian in origin.

The Central German, or Franconian, dialect and the Thuringian dialect helped to form the basis of modern standard German. The present-day influence of Thuringian is of greatest significance in Thuringia, Saxony, and Saxony-Anhalt states. East Franconian is spoken in northern Bavaria, South Franconian in northern Baden-Württemberg. The Rhenish Franconian dialect extends northwest from approximately Metz, in French Lorraine, through the states of Rhineland-Palatinate and Hessen. Moselle Franconian extends from Luxembourg through the Moselle valley districts and across the Rhine into the Westerwald. Ripuarian Franconian begins roughly near Aachen, at the Dutch-Belgian border, and spreads across the Rhine between Düsseldorf and Bonn into the Sauerland.

The dialect known as Low German, or Plattdeutsch, historically was spoken in all regions occupied by the Saxons and spread across the whole of the North German Plain. Although it has been largely displaced by standard German, it is still widely spoken, especially among elderly and rural inhabitants in the areas near the North and Baltic seas, and is used in some radio broadcasts, newspapers, and educational programs. Tiny pockets of Frisian, the German dialect most closely related to English, persist. Foreign immigration, more widespread education, the influence of the United States, and globalization also have helped create a polyglot of languages in major German cities.

RELIGIONS

The Reformation initiated by Martin Luther in 1517 divided German Christians between Roman Catholicism and Protestantism. The Peace of Augsburg (1555) introduced the principle that (with some exceptions) the inhabitants of each of Germany's numerous territories should follow the religion of the ruler; thus, the south and west became mainly Roman Catholic, the north and east Protestant. Religious affiliation had great effect not only on subjective factors such as culture and personal attitudes but also on social and economic developments. For example, the willingness of Berlin to receive Calvinist religious refugees (Huguenots) from Louis XIV's France meant that by the end of the 17th century one-fifth of the city's inhabitants were of French extraction. The Huguenots introduced numerous new branches of manufacture to the city and strongly influenced administration, the army, the advancement of science, education, and fashion. The Berlin dialect still employs many terms of French derivation.

Population movements during and after World War II brought many Protestants into western Germany, evening the numbers of adherents of the two religions. In the former West Germany most people, whether or not they attended church, agreed to pay the church tax levied with their income tax; the revenue from this tax has been used to support community centres, hospitals, senior citizens' centres and group homes, and the construction of church buildings in the former East Germany. The centrality of religion in Germany has meant that religious leaders, especially the Roman Catholic hierarchy, sometimes exercise considerable influence on political decisions on social issues such as abortion.

In East Germany Protestants outnumbered Roman Catholics about seven to one. Although the constitution nominally guaranteed religious freedom, religious affiliation was discouraged. Church membership, especially for individuals who were not members of the ruling Socialist Unity Party (SED), was a barrier to career advancement. Similarly, youth who on religious grounds did not join the Free German Youth (Freie Deutsche Jugend) lost access to recreational facilities and organized holidays and found it difficult, if not impossible, to secure admission to universities. Not surprisingly, formal church affiliation was relatively low, amounting to only about half the population, compared with nearly seven-eighths in West Germany. However, Protestant (Lutheran) churches did act as rallying points for supporters of unofficial protest groups, leading ultimately to the demonstrations that toppled the communist government in 1989.

Lutherans and Roman Catholics in Germany now are about equal in number. Small percentages of Germans are members of what are known as the free churches, such as Evangelical Methodists, Calvinists, Old Catholics, Jehovah's Witnesses, and (by far the largest) Eastern Orthodox. The number of people professing no religion (*Konfessionslose*) has sharply increased and now represents about one-fifth of all Germans. Because of large-scale Turkish immigration, Muslims now account for some 5 percent of the total population. Only a few thousand German Jews survived the Holocaust. During the 1990s, however, Germany's Jewish population quadrupled, the result of significant immigration from eastern Europe (especially Russia). There are now some 100,000 Jews in the country, and Berlin, with Germany's largest concentration of Jews, has experienced a modest rebirth of its once thriving Jewish community.

Konfessionslose

DEMOGRAPHIC TRENDS

"Internal" migration. After World War II Germany received more than 12 million refugees and expellees from former German territory east of the Oder and from areas with substantial German ethnic populations in central and eastern Europe. These numbers were swollen by the ranks of "displaced persons"—non-Germans unwilling to return to their former homelands. After Germany was partitioned in 1949, the demographic histories of the two parts of the country diverged, with West Germany becoming the prime target of continuing migration flows. Although immigrants, principally ethnic Germans, continued to drift in from the east, their numbers were overshadowed by a mass desertion of some two million people from East Germany. Because these immigrants from East Germany were mostly young and highly skilled, their arrival was a major gain for the booming West German economy but a grievous loss for the much smaller East Germany. In 1961 the East German government blocked further flight by its people by building strong defenses along the inner-German border and around West Berlin (including the Berlin Wall). East Germany enjoyed relative demographic tranquillity for most of the following three decades. After the disintegration of communist regimes throughout central and eastern Europe, however, the population of West Germany began to surge again, because of flows first from newly liberalized Hungary and Czechoslovakia and then from East Germany after the inner-German boundary was opened and the Berlin Wall fell in November 1989. In 1989–90 alone

nearly 700,000 East Germans poured into West Germany; thereafter the stream continued, though from 1994 to 1997 net immigration occurred at a sharply reduced rate before increasing again because of ongoing economic problems in eastern Germany.

The arrival of these new migrants caused some resentment among western Germans because of the pressures placed on an already overburdened housing market and on social services. Because the new arrivals were mainly young and skilled, they fueled a postunification boom in western Germany but continued to drain the economy and society in the east, which still faces economic and social problems. Several hundred thousand eastern Germans also commuted to jobs in western Germany.

"Guest workers"

Immigration. To spur economic growth, West Germany began as early as the mid-1950s to encourage workers to migrate from other countries. At first these migrants were to be "guest workers," coming to work for a limited period of time only, but increasingly they sent for their families; thus, even when economic recession occurred in 1973 and the further immigration of workers was discouraged, the number of foreign residents continued to grow, reaching more than seven million people—nearly one-tenth of the total—by the beginning of the 21st century. Because of higher birth rates among the foreign-born population, non-Germans have accounted for a majority of natural population growth since the 1950s. Turks represent the largest group of foreign residents, followed by Serbs and Montenegrins, Italians, Greeks, Poles, Croats, Austrians, and Bosnians. Immigrants typically were employed in the heaviest, dirtiest, and least-remunerative jobs, and in times of economic difficulty they generally were the first to lose their jobs and the last to be reemployed. Their children—of whom more than four-fifths have been born in Germany—are among the last to be considered for an apprenticeship or training place. Immigrants also inhabit the least-desirable housing. Turks, in particular, have formed distinctive quarters in the poorest "inner city" areas. Although the East German state prided itself on its nonreliance on guest workers, some Poles, Vietnamese, Angolans, Cubans, and Mozambicans were imported, ostensibly for "education and training."

With the opening of the eastern frontiers and a more liberal attitude of the Soviet Union toward emigration, the influx of ethnic Germans from former Soviet-bloc countries other than East Germany became a veritable flood. Nearly 400,000 came in 1989, followed by more than 200,000 annually between 1991 and 1995; subsequently the number of immigrants fell but remained substantial. These new immigrants were less easily assimilated into western German culture than those from eastern Germany; many had difficulties with the German language and lacked marketable skills. With some apprehension, united Germany realized that an additional one million ethnic Germans could arrive from eastern Europe in the future, and there was a further fear that the freedom to travel and political or economic problems might produce a flow of untold millions of non-German residents of the former Soviet Union. Partially in response to these concerns, Germany's relations with Russia focused on attempting to improve the lot of ethnic Germans living in Russia, thereby diminishing the likelihood of mass emigration to Germany.

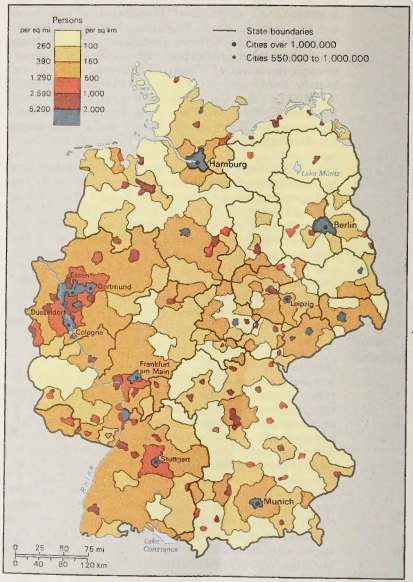
The right of asylum

West Germany's constitution guaranteed the right of asylum to those forced to flee their native countries because of political oppression. This privilege was regarded as compensation for the asylum granted to 800,000 German victims of political and ethnic persecution during World War II. Criticism of this constitutional provision mounted in the 1980s with the arrival of asylum seekers from non-European countries such as Sri Lanka, Iran, Lebanon, Ghana, and India, together with stateless Palestinians; it was difficult to distinguish those hoping to better themselves economically or to avoid compulsory military service from genuine victims of oppression. The issue of asylum became even more pressing when the eastern borders were opened, admitting a flood of foreigners—most prominently Poles, Romanian Roma (Gypsies), and Bosniacs (Bosnian Muslims). Between 1990 and 1993, one million people sought asylum in Germany, and, as antagonism toward immigrants increased, there was a surge of violent attacks against foreigners. Although the government and citizen groups condemned such xenophobic sentiment and behaviour, foreigners continued to be subjected to discrimination and sporadic violence. Beginning in 1991, legislation brought Germany in line with the more restrictive policies practiced by other members of the European Community (EC; since 1993 embedded within the European Union) regarding immigration from outside the EC. But while cooperation with neighbouring states reduced the flow of illegal immigrants and somewhat abated the problem, Germany nevertheless became embroiled in a domestic debate over the rights of noncitizen residents, including the right to naturalization, which had become somewhat easier for long-term residents in the late 1990s.

POPULATION STRUCTURE

General characteristics. Germany is the most populous European country west of Russia. Its population density is high in comparison with most other European countries, though it is exceeded by those of Belgium and The Netherlands. Germany has one of the world's lowest birth rates, and its life expectancy—some 75 years for males and 80 for females—is among the world's highest. Over the last several decades Germany has witnessed years of both positive and negative population growth. From the mid-1970s to the mid-1980s the country's population dropped; however, Germany experienced significant population growth—largely because of immigration—over the following decade. Thereafter the country's population growth was slight. To stem long-term population decline, governments at all levels have attempted to develop policies aimed at encouraging an increase in the birth rate, in particular by subsidizing child care and providing benefits and other tax incentives to families.

As in most industrialized countries, the proportion of the



Population density of Germany.

population under age 15 is quite low, accounting for about one-sixth of the total; in contrast, the proportion of those over age 60 has increased dramatically, representing more than one-sixth of the population. That women predominate in the older population is largely a reflection of their higher life expectancy and of the significant losses of men during World War II, but the discrepancy has become less pronounced with time. In eastern Germany the disproportion of elderly citizens increased significantly after the desertion of the young before the erection of the Berlin Wall and after its destruction in 1989.

Population distribution. Because Germany has for centuries had a profusion of states (each with towns and cities), its population is more widely dispersed than that of countries, such as France, in which centralization occurred early. It is possible, however, to discern two major population axes. The main axis runs from the Rhine-Ruhr region southward through the Rhine-Main (Frankfurt) and Rhine-Neckar (Heidelberg-Mannheim-Ludwigshafen) agglomerations, the great cities of southern Germany, and to Basel (Switzerland) and the Alpine passes. This is the main axis not only for Germany but also the European Union (EU). The second axis runs from the Rhine-Ruhr region eastward north of the Central German Uplands through Hannover, Braunschweig, and Magdeburg to the great urban concentration of Saxony. Some major cities stand in isolation outside the two axes, notably Augsburg and Nürnberg to the south and Bremen, Hamburg, and the capital city of Berlin, forming islands in the thinly populated North German Plain. Before unification, population redistribution from the agglomeration cores in western Germany was accompanied by a marked drift of population from the north to the booming cities and attractive environment of southern Germany. In eastern Germany, early gains by migration were experienced in areas of planned industrial development (Eisenhüttenstadt, Rostock, Schwedt, Hoyerswerda). After initial postwar recovery, the cities of the south lost population, reflecting industrial decline, rationalization of production, and unattractive environments. East Berlin and its satellite towns were the principal targets of migration until western Germany became accessible in 1989.

Population axes

The economy

The German constitution, the Basic Law, guarantees the right to own property, freedom of movement, free choice of occupation, freedom of association, and equality before the law. However, the constitution modified the operation of the unfettered free market by means of its "social market economy" (*Soziale Marktwirtschaft*). With a "safety net" of benefits—including health care, unemployment and disability compensation, maternity and child-care provisions, job retraining, pensions, and many others—paid for by contributions from individuals, employers, and public funds, Germany has an economic order supported by most workers and businesses.

In the social market economy, the government attempts to foster fair play between management and labour and to regulate the relationship between the capitalist participants in the market, particularly with regard to competition and monopolies. Works councils have been established, and workers have representation on the boards of businesses. The social market economy was created by policy makers with a vivid memory of market distortions and social tensions caused by the giant industrial trusts before 1939. Legislation against monopolies appeared in 1958 and has been criticized as ineffective. For example, it has proved impossible to restrict the indirect coordination, through which individuals, banks, and other financial institutions build up "diagonal" share holdings linking a range of firms that are nominally independent. Moreover, where a whole branch of industry has experienced difficulties (*e.g.*, the Ruhr coal industry), even the federal government has encouraged concentration. The emergence of very large monopolistic firms has been unavoidable because, in an increasingly international economy, large firms that enjoy economies of scale are better positioned to survive. With globalization, governments are less able to regulate businesses at the national level or even at the transnational level of the EU.

Social market economy

The social market economy is regulated and administrated not exclusively by the federal government but by a plurality of agencies. For example, there are numerous insurance institutions that deliver social benefits. The most important institution in post-World War II Germany is the Frankfurt-based Deutsche Bundesbank (German Federal Bank). With memories of the runaway inflation of 1922–23, the West German government decided that it should never again have a license to print money and that the central bank should be independent of political control. Consequently, Germany's adoption of the euro, the EU's single currency, in 1999 raised some concerns in the country that the European Central Bank would be subject to political influence and manipulation. The Chambers of Trade, at every level of the administrative hierarchy, are also influential, and the state governments play a significant economic role (*e.g.*, the government of North Rhine–Westphalia is intimately concerned with the survival of the Ruhr coal industry). Federal and state governments also participate in the ownership of some enterprises, notably public utilities. The Basic Law, however, prevents the arbitrary intervention of the central government.

Because Germany has numerous economic actors, a high degree of coordination has been required to achieve adequate growth, balanced foreign trade, stable prices, and low unemployment. A variety of consultative bodies unite federal and state governments, the Deutsche Bundesbank, representatives of business and of the municipalities, and trade unions. The Board of Experts for the Assessment of Overall Economic Trends, established in 1963 and known as the "five wise men," produces an evaluation of overall economic developments each year to assist in national economic decision making. Moreover, the federal government submits an annual economic report to the legislature that contains a response to the annual evaluation of the Board of Experts and an outline of the economic and financial policies it is pursuing.

Although the free market operates in Germany, the federal government plays an important role in the economy. It is accepted as self-evident that it should underwrite the capital and operating costs of the economic and social infrastructure, such as the autobahn network, waterways, the postal system and telecommunications, and the rail system. The federal government, the states, and the cities also contribute to the regional and local rapid transit systems. Government collaborates with industry in bearing the costs of research and development, as, for example, in the nuclear power industry. Federal intervention is particularly strong in the defense industry. The coal industry is perhaps the most notable example of subsidization, and agriculture has traditionally been massively protected by the state, though the sector is now governed by EU institutions. Regional planning is another significant field of government intervention; the federal government fosters economic developments in rural and industrial "problem" regions. States and cities also intervene with schemes to foster regional or local development.

Germany has a varied tax system, with taxes imposed at the national, state, and local levels. Because of the generous system of social services, tax rates on corporations, individuals, and goods and services are all relatively high in comparison with those of other countries. Germany employs a system of tax equalization, through which tax revenues are distributed from wealthier regions to less-prosperous ones. After unification these transfers were resented among many western Germans.

FROM PARTITION TO REUNIFICATION

The West German system. After the devastation of World War II, West Germany rebounded with a so-called "economic miracle" that began in 1948. The subsequent combination of growth and stability made West Germany's economic system one of the most respected in the world, though it began to suffer strains beginning in the 1990s, exacerbated by the costs of unification. Germany's remarkable economic performance was largely a result of

Postwar economic miracle

effective economic management, but temporary factors were especially important in spurring economic growth in the immediate post-World War II era. In particular, a large force of unemployed workers—returned servicemen and displaced persons—were available and eager to rebuild their own lives and willing to work hard at a rate of remuneration that left a considerable investment surplus in their employers' hands. In addition, the country reaped benefits from the joint economic planning for the American, British, and French zones of occupation that culminated in the vital and essential currency reform that introduced the deutsche mark in June 1948 and the U.S.-financed Marshall Plan (1948–52), which helped to rebuild war-torn Europe.

From 1951 to 1961 West Germany's gross national product (GNP) rose by 8 percent per year—double the rate for Britain and the United States and nearly double that of France—and exports trebled. Despite some occasional economic downturns (*e.g.,* during the oil crisis of 1973–74), West Germany's economy followed an upward trend. Indeed, when East and West Germany reunited in 1990, West Germany's economy was enjoying a cycle of business expansion that had begun in the early 1980s and continued into 1992. By that time Germany had one of the largest economies in the world and was a leader in world trade. All this was achieved while maintaining low inflation.

The East German system. East Germany also had experienced an economic miracle of sorts. Unlike the other Soviet-style states of eastern Europe, East Germany had been part of an advanced capitalist economy before the war, which gave it a considerable advantage in reconstruction. Even though it had emerged from World War II and the postwar Soviet demolitions economically ravaged, its surviving industrial infrastructure, inherited skills, and high level of scientific and technical education enabled it to develop the economy and to advance the standard of living to a level markedly higher than those of most other socialist countries, though living standards were still well below those of western Europe. East Germany became the principal supplier of advanced industrial equipment to the communist countries, though it became apparent after unification that it produced poor quality goods and caused environmental devastation.

East Germany had a command economy, in which virtually all decisions were made by the governing communist party, the Socialist Unity Party (SED). The system of planning was inflexible and eventually caused ruinous economic conditions. Power, influence, and personal connections (*Beziehungen,* or "vitamin B") drove economic decisions, and all groups, including trade unions, were expected to collaborate to achieve the SED's economic objectives.

"Vitamin B"

East Germany's industrial sector lacked quality controls and technological innovation. The cynicism, apathy, and inertia that were common among workers and enterprise managers contributed to low rates of East German technological development. Despite excellent training, workers were not rewarded with increased earnings for ingenuity; the result was a general malaise.

Supply and distribution were controlled by state-owned companies, and the centralized provision of services through nationalized concerns and local administrations was a generally recognized weakness. This was partially addressed by a "gray market" for goods and services in short supply (*e.g.,* automobiles and automobile and house repairs), particularly when payment was made in hard currency; for example, repairmen offered much faster service for an extra fee or favours, and sales clerks also kept certain goods "under the counter." By the 1970s and '80s, particularly as contacts with the West increased, this gray market grew in significance.

ECONOMIC UNIFICATION AND BEYOND

The implementation of Mikhail Gorbachev's glasnost (political liberalization) and perestroika (economic restructuring) policies in the Soviet Union fueled sentiment in Germany that reunification could become a reality, and the basic steps toward German economic unity were accomplished with astonishing speed. The unexpected opening of the frontier between East and West Germany and the breaching of the Berlin Wall on Nov. 9, 1989, were a heavy blow to the East German economy, as the relatively small numbers of migrants, who in previous years had left the country by way of Hungary or Czechoslovakia, rose dramatically. Exacerbating the problem was the fact that most of those who left were the younger, more active members of the population and those with marketable skills. The economic unification, achieved by July 1, 1990, swept away all customs barriers and introduced the deutsche mark as the sole currency in Germany.

Following Germany's official reunification on Oct. 3, 1990, the western German economy continued to grow rapidly until 1992, after which it began to experience an economic slowdown before growth resumed in the mid-1990s. During the decade following 1992, the German economy grew at an average annual rate of 1.4 percent—among the lowest rates in western Europe. Many economists attributed the slowdown to rigid labour policies, high taxes, marginal incentives for investment, and generous incentives for workers to retire, miss work, or be unemployed. The slowdown was also related to unification, which wholly revealed the economic deficiencies of East Germany—the extent of its technological backwardness, its low productivity, and the faltering state of its manufacturing plants. Disillusionment in eastern Germany rose sharply as manufacturing output and employment declined rapidly. The federal government's insistence that eastern German firms compete immediately in the free market led to economic devastation in the east. By spring 1991, mass demonstrations against unemployment occurred regularly in Leipzig, and there was concern that economic despair would cultivate the rise of political extremism. Indeed, the Berlin office of the Treuhandanstalt (a government-owned but independent trust agency for the privatization of eastern German industry with wide powers of disposal) was firebombed, and in April 1991 its head was murdered by the West German Red Army Faction.

Low growth rates

The Deutsche Bundesbank believed that the government had introduced the deutsche mark into eastern Germany too precipitately, with practically no preparation or possibility of adjustment, and at too favourable a rate. The effect of currency conversion and subsequent wage pressure deprived industry in the east of one of its few advantages, low labour costs. The favourable exchange rate and relatively high wages and salaries did, on the other hand, help achieve a sociopolitical goal—encouraging eastern Germans to remain in the east rather than migrating to the west, where people feared being overwhelmed by migrants. There were commercial bankruptcies in eastern Germany, and the eastern economy was further decimated by the tendency of easterners to buy the better-presented and technically superior consumer goods from western Germany or abroad rather than the generally drab products of eastern German industry; by the end of the decade, however, high-quality goods produced in eastern Germany bolstered the economy, and there was a wave of regional consciousness that favoured the patronage of local products.

Economic unification caused particularly severe hardships for eastern German workers; unemployment rose sharply and industrial output fell by two-thirds in the years after unification. Decline was greatest in the food-processing sector, metallurgy, building materials, machinery and vehicles, electronics and related equipment, and textiles. Eastern German agriculture also was devastated, with employment dropping by some three-fourths. Although the eastern economy later rebounded, at the beginning of the 21st century more than one-sixth of its labour force was unemployed—more than double the rate for western Germany. Unemployment also rose disproportionately for women. As a result of job losses, migration from east to west continued throughout the 1990s and into the early 21st century.

Unemployment in eastern Germany

The slowness of economic recovery in eastern Germany was the result of a variety of factors. The haste of change, especially regarding the currency conversion and the breakup of the great industrial combines, and the fact that East Germany had no effective government for a period of

three months following the economic union in July 1990 hampered economic reconstruction efforts. Even after political unification, progress was disappointing. Firms removed from ministerial control and transformed into limited companies found themselves unable to compete in the free market, burdened not only with outdated plants but with debt, because the East German government had appropriated their profits while requiring them to borrow their capital. The federal government had assumed that the reconstruction of eastern German industry would essentially come about by the takeover of plants by Western, predominantly western German, firms. In reality, however, the Treuhandanstalt set up to dispose of some 10,000 formerly nationalized firms made extremely slow progress, partly as a result of an excessively legalistic approach and partly because of the shortage of experienced administrators afflicting the reconstituted public service in the east. Western German firms were under no great financial pressure to move in, and, with the help of the additional labour available from eastern German migrants, they expanded production at their existing plants without having to become involved in the difficulties of actually setting up a branch in the east. Protesters warned that eastern Germany was turning into an internal colony; however, this overly pessimistic outlook was exaggerated, and about 1992 some economic revival began to occur.

Land ownership was a significant barrier to establishing plants in eastern Germany. Following the principles of the German constitution, after unification, former owners were assured that they could repossess their property or at least be compensated for their losses. However, this did not apply to property expropriated by the Soviet military administration (1945–47), including many large estates that not everybody would be happy to see returned to their original aristocratic owners. Where a plant had originally been owned by a family or firm in western Germany but had received additional investment from the East German government and had perhaps expanded over land originally in a number of hands, western German firms were deterred from moving in, there being a lack of clear title to ownership.

The production-focused East German communist system had ignored environmental considerations. Firms seeking to take over electrical generation based on brown coal, any part of the chemical industry, or any other plant where dangerous chemicals had been used in processing faced enormous costs in attempting to meet federal government standards. Firms were also discouraged from taking over plants, because the inevitable reductions in surplus labour would involve the payment of unemployment compensation. As a result, the few western German firms setting up in the east preferred to establish a completely new plant on a green-field site, allowing them to avoid these excessive costs.

The federal government initially believed that the costs of unification could be borne by borrowing and without increases in taxation. Despite these assurances by Chancellor Helmut Kohl during the 1990 all-German election campaign, by 1991 additional taxation was required. If people in the east were disillusioned by the economic results of union, those in the west grew increasingly resentful of the cost of paying for it.

Costs of unification

During the 1990s Germany made a number of dramatic changes in its energy sector (*e.g.*, higher taxes, lower subsidies for coal mining, and privatization of huge eastern German energy firms), and in 2000 the government announced a plan to phase out the nuclear power industry by about 2025. Massive reconstruction projects in the east (*Aufbau Ost*), funded largely by higher taxes in the west, helped to improve infrastructure in the eastern regions. Telecommunications systems were upgraded, and there were generous subsidies to encourage capital investment.

Quite apart from the costs and problems associated with unification, Germany and its economy faced a number of interrelated problems at the beginning of the 21st century. High unemployment—which regularly exceeded four million people—became the chief political issue. Extremely high wages—among the world's highest—generous social services, and high taxation also dampened the economy. Unification caused the public debt to grow dramatically, and at the beginning of the 21st century some one-fifth of the annual federal budget went toward interest payments on the accrued national debt.

Although unification was more than a decade old, at the beginning of the 21st century its effects still weighed heavily on the German economy and its political institutions. However, in large measure unification gave way to other issues, such as globalization, the introduction of the euro as the single currency of the EU in 2002, and the enlargement of the EU to include central and eastern European countries. Germany's domestic economic problems and opportunities are complexly bound up with global and regional processes over which it has only varying levels of influence and control—a somewhat unsettling situation for a society that became very prosperous by following accustomed patterns and having firm control of the major levers of its own economy.

AGRICULTURE, FORESTRY, AND FISHING

Agriculture. As in other sectors of the economy, the division of Germany was reflected in a dramatic divergence of agricultural development. West Germany remained essentially a country of small family farms; in the 1980s only about 5 percent of holdings had more than 124 acres (50 hectares), though they accounted for nearly one-fourth of the total agricultural area. By the beginning of the 21st century, however, large farms represented about half of the total agricultural area in western Germany and some two-thirds in eastern Germany. The change in western Germany is reflective of a rationalization of agriculture, with many small landholders leaving farming and the remaining farms often increasing in size. The larger farms in the west are mainly concentrated in Schleswig-Holstein and eastern Lower Saxony, with smaller groupings in Westphalia, the lowland west of Cologne, and southern Bavaria. Small farms predominated in the central and southern parts of West Germany. The process of steady enlargement decreased the total number of holdings by more than three-fourths from 1950 to the end of the 20th century. The number of people employed in agriculture also declined substantially, from about one-fifth of the total workforce in 1950 to less than 3 percent by the end of the 20th century. Wage labourers virtually disappeared from all but the largest farms, and smaller farms were cultivated on a part-time basis.

By contrast, in the east, following conquest by the Soviet army at the end of World War II, many large estates were split up or retained as state farms. From 1952 to 1960 virtually all the small farms in East Germany were united, under strong political pressure, to form agricultural cooperatives. Agricultural production was increasingly concentrated into extremely large specialized units; by the mid-1980s state-run or cooperative crop-producing enterprises averaged more than 11,000 acres (4,450 hectares). Despite a marked decrease in agricultural workers, modern machinery and technological innovation led to increased production. After unification agricultural employment in eastern Germany plunged by about three-fourths.

The agricultural cooperatives

In areas of high natural fertility, wheat, barley, corn (maize), and sugar beets are the principal crops. The poorer soils of the North German Plain and of the Central German Uplands are traditionally used for growing rye, oats, potatoes, and fodder beets. Technological changes have altered much of the traditional spatial pattern of German agriculture. Sugar beets, formerly confined to deep fertile soils such as the loess lands on the northern fringe of the Central German Uplands, are now much more widespread. With the availability of chemical fertilizers, light soils have become more highly valued because of their suitability for machine cultivation; for example, fodder corn is now widely grown on the North German Plain, replacing potatoes. The two most widespread forms of agricultural land use are cereal cultivation (including corn for its grains) and permanent pasture; both are important sources of animal feed. Dairying formerly was concentrated in the area of mild climate in the northern coastal lowlands and in the Alpine foothills, but it is now widespread in all areas where small farms predominate. East Germany concen-

trated milk production into vast specialist holdings in arable areas where food was available and urban markets accessible. In both the western and eastern sectors, chickens, eggs, pigs, and veal calves are concentrated into large battery units, divorced from immediate contact with the soil. Besides concern for the plight of the animals under this system of concentrated production, Germans are distressed by the groundwater pollution associated with it.

In the areas surrounding western German cities, crops such as fruits, vegetables, and flowers are grown. The warm lowlands of the southwest favour tobacco and seed corn. They also support vegetables, as do the Elbe marshes south of Hamburg and the marshy Spreewald south of Berlin. Fruit grows abundantly in southern Germany; other important areas of specialization include the "Altes Land" on the Elbe south of Hamburg, the Havel lake country near Potsdam, and the Halle area. Vineyards are located in the west, especially in or near the valleys of the Rhine, Moselle, Saar, Main, and Neckar rivers, although the slopes of the Elbe valley near Dresden also produce wine grapes.

Vineyards

At the time of reunification, western Germany produced some four-fifths of its food requirements, and increased productivity and guaranteed prices resulted in vast surpluses (especially of butter, meat, wheat, and wine). At the beginning of the 21st century, Germany's production of major agricultural products (*e.g.*, grains, sugar, oils, milk and meat) exceeded domestic consumption, resulting in both exports and continued surpluses.

Forestry. Some three-tenths of Germany's total land area is covered with forest. In the Central German Uplands and the Alps, forests are particularly plentiful, but they are notably absent from the best agricultural land, such as the loess areas of the North German Plain. The western part of the North German Plain also has little forest cover, but there are substantial wooded stretches farther east. Conifers predominate in the forest area; spruce now accounts for much of the plantings because of its rapid growth and suitability for building purposes and for the production of paper and chipboard. Domestic production covers about half of the demand for wood from temperate forests, but producers face severe competition from Austria, Scandinavia, and eastern Europe. The federal government, states, and municipalities own about half the forest in western Germany, with the remainder in private hands; eastern German forests are primarily publicly owned.

Fishing. Fishing in western Germany began to decline markedly from the 1970s because of overutilization of traditional fishing grounds and the extension of the exclusive economic zone to 200 miles (320 kilometres) offshore. The greatly reduced deep-sea fleet now uses freezer vessels and accompanying catchers; Bremerhaven, Cuxhaven, and Hamburg are the home ports and processing centres. During the 1990s, high-seas catches by German fishermen declined by about half. The North Sea herring fishery has almost disappeared, and now the German appetite for pickled herring is satisfied mainly by imports. There are well over 100 fishing ports on the North Sea and Baltic coasts. Fishing for shrimp and mussels is important on the mud flats fringing the North Sea. Prior to unification East Germany had a substantial deep-sea fishing fleet, but most of it has since been scrapped; its shore base for fish processing was at Sassnitz on the island of Rügen.

RESOURCES AND POWER

Germany, which has relatively few domestic natural resources, imports most of its raw materials. It is a major producer of bituminous coal and brown coal (lignite), the principal fields of the latter being west of Cologne, east of Halle, south and southwest of Leipzig, and in Lower Lusatia in Brandenburg. Other minerals found in abundance are salt and potash, mined at the periphery of the Harz mountains. The mining of most metallic minerals ceased for economic reasons in western Germany before unification; in the 1990s the centuries-old mining and processing of copper ores in the Mansfeld area of eastern Germany and the mining and processing of uranium ores for the benefit of the Soviet Union in the Ore Mountains also stopped. There are small reserves of oil and natural gas in northern Germany.

Coal

As in all industrialized countries, water supply is a constant problem. The filtration of water on riverbanks (*e.g.*, those of the Rhine) is one source. It is supplemented by reservoirs in the uplands. For example, the Harz mountain range provides water to much of the North German Plain as far as Bremen, and the Ore Mountains supply the central German industrial region.

Oil is Germany's principal source of energy. As domestic production is quite limited, most crude oil is imported. Many petroleum products also are imported, transported from Rotterdam by product lines, barges, and rail. Until the mid-1950s the refining of oil took place at the coast, notably at Hamburg and Rotterdam; however, refineries have been developed at inland locations close to markets, mostly on rivers such as the Rhine and Danube, which are served by pipelines from Wilhelmshaven, Rotterdam (Netherlands), Lavéra (near Marseille, France), Genoa (Italy), and Trieste (Italy). Eastern Germany receives oil delivered by pipeline from Russia to a refinery at Schwedt on the Oder, which supplies the central German industrial region; there is also a pipeline from Rostock that provides industry with oil. German supplies of natural gas are significant, but most gas is imported. Principal sources are the Friesian and North Sea fields of The Netherlands and the Norwegian North Sea. Gas is imported from Russia via a pipeline from the Czech Republic, with a branch serving eastern Germany and Berlin.

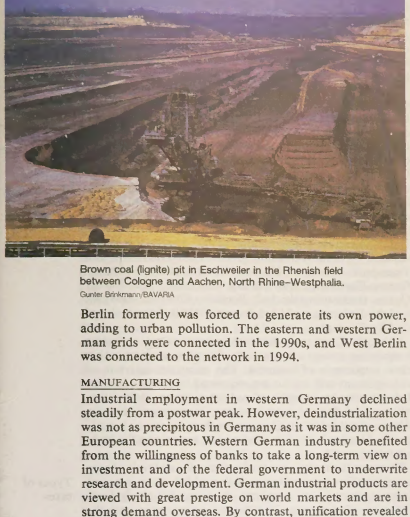
Bituminous coal, Germany's second most important source of energy, is available in profusion from the Ruhr field and from the smaller Saar, Aachen, and Ibbenbüren fields, though extraction is costly and often subsidized. In the last half of the 20th century, however, output shrank by some two-thirds. Coal now has two major uses: the generation of electricity and the production of metallurgical coke. A striking feature of the German economy is the significance of brown coal (lignite). This low-grade, waterlogged fuel can be worked economically in vast open pits, which are mined with massive machines. About seven-eighths of all the coal is fed straight to electric-power generating stations that are situated on the field itself. A relatively small quantity of the coal is pressed into briquettes for domestic heating. Electricity generation is also the principal use of the main fields in eastern Germany; however, during partition lignite was a major basis of the chemical industry as well as a source of gas and briquettes for urban consumption. After unification many eastern German pits closed, particularly those producing the most sulfurous coal. The shortfall in energy output led the federal government to subsidize additional imports of gas from Russia.

The largest producers of electric energy are the thermal plants that are located primarily in the Ruhr and the Rhenish brown-coal fields and in the brown-coal fields of the east, especially in Lower Lusatia. During partition all western German plants were required to significantly reduce the emissions of the dust, sulfur dioxide, and nitrogen oxide formerly emitted into the atmosphere. Plants in the east were not similarly regulated and thus contributed to general atmospheric pollution; after unification a number of them were closed and others were upgraded.

Nuclear power plants rival thermal plants in significance; in western Germany they are typically located on the coast or on rivers far from the coalfields. Plants in eastern Germany, built on the Soviet (Chernobyl) model, were closed for safety reasons. In 2000 the German government committed to phasing out all of the country's nuclear power plants within about 20 years, though the future of nuclear power, as well as the number of years that existing plants would function, was unclear.

Nuclear power

The canalization of such rivers as the Main, Neckar, and Moselle, together with hydroelectric power plants in the Alps, produce relatively minor amounts of electric power; pumped storage schemes in mountain areas are important in meeting peak electricity demands. Before unification, East and West Germany had distinct transmission grids without interconnection. The West German network was linked to that of neighbouring countries, allowing it to import surplus power from the French nuclear system and, during the Alpine snow melt, especially from Austria. West



Brown coal (lignite) pit in Eschweiler in the Rhenish field between Cologne and Aachen, North Rhine–Westphalia.
Gunter Brinkmann/BAVARIA

Berlin formerly was forced to generate its own power, adding to urban pollution. The eastern and western German grids were connected in the 1990s, and West Berlin was connected to the network in 1994.

MANUFACTURING

Industrial employment in western Germany declined steadily from a postwar peak. However, deindustrialization was not as precipitous in Germany as it was in some other European countries. Western German industry benefited from the willingness of banks to take a long-term view on investment and of the federal government to underwrite research and development. German industrial products are viewed with great prestige on world markets and are in strong demand overseas. By contrast, unification revealed that most of eastern German industry was incapable of competing in a free market.

Steel manufacture

Germany is one of the world's leading manufacturers of steel, with production concentrated in the Ruhr region; however, since the peak output of the early 1970s, a number of plants have closed. (The steel industry in eastern Germany was largely abandoned after unification, though some production was reestablished at a renovated plant at Eisenhuettenstadt.) Germany's principal industries include machine building, automobiles, electrical engineering and electronics, chemicals, and food processing. Automobile manufacturing is concentrated in Baden-Württemberg, Lower Saxony, Hessen, North Rhine–Westphalia, Bavaria, the Saarland, and Thuringia. Leading automobile manufacturers in Germany include Audi, BMW, Daimler-Chrysler (formerly Daimler-Benz), Ford, Opel, and Volkswagen. Following unification, production of the environmentally unfriendly Trabant and Wartburg cars in eastern Germany ceased. Volkswagen, Opel, and Daimler-Benz were quick to establish assembly or parts production in the east. Shipbuilding, once a major industry, has declined significantly.

Since the late 19th century Germany has been a world leader in the manufacture of electrical equipment. As the home of internationally known firms such as Siemens, AEG, Telefunken, and Osram, Berlin was the industry's principal centre until World War II, after which production was largely transferred to Nürnberg-Erlangen, Munich, Stuttgart, and other cities in southern Germany. The output of these centres made Germany one of the world's leading exporters of electrical and electronic equipment.

In East Germany electrical and electronic production was concentrated in East Berlin, with Dresden forming a second important centre. The country was a major supplier of electronic equipment (*e.g.*, computer-controlled robots) to the communist world. Although eastern German plants were outdated in comparison with those in the west, both Dresden and Erfurt achieved some success in developing microelectronics production following unification.

With the discovery of synthetic dyestuffs in the late 19th century, Germany became a world leader in the chemical industry. Most of the western German chemical industry is concentrated along the Rhine or its tributaries, notably in Ludwigshafen, Hoechst (near Frankfurt), and Leverkusen (together with a row of other plants along the Rhine in North Rhine–Westphalia). Chemical plants also operate in the Ruhr region. The majority of East German chemical plants were on the two brown-coal fields of Lower Lusatia and Halle-Leipzig; after unification some plants were closed because of environmental reasons, and others were upgraded.

Germany is also particularly strong in the field of optical and precision industries. The once-mighty textile industry has suffered from overseas competition but is still significant. Principal centres are in North Rhine–Westphalia (Mönchen-Gladbach, Wuppertal) and southern Germany. After unification many textile plants were closed in eastern Germany, where employment in the sector plunged by some nine-tenths.

FINANCE

The central banking system. Germany's central bank, the Deutsche Bundesbank, is headquartered in Frankfurt am Main, which is the country's main financial centre and also the base of the European Central Bank, the EU's chief financial institution. Before the circulation of the euro, the common currency of the EU, in 2002, the Bundesbank issued the deutsche mark (the country's former currency) and oversaw its circulation. As the EU's most powerful national central bank, the Bundesbank played a pivotal role in the planning of and preparation for the euro. One of its primary roles now is to implement the monetary policies of the European System of Central Banks to help maintain the euro's stability.

The role of the Bundesbank

Upon the establishment of the Bundesbank, its preeminent characteristic was its independence from government control, instituted to prevent a recurrence of the severe inflation experienced in 1922–23, when the government resorted to the printing press for finance. The federal bank maintained a policy of careful control of credit and concern for the international exchange rate of the deutsche mark, which had made West Germany the leading financial power in post-World War II Europe. The Bundesbank demonstrated its genuine independence in 1991 when it insisted that additional government expenditure for the eastern sector be covered by unwelcome tax increases rather than by borrowing. Individual *Land* (state) central banks are the Bundesbank's representatives at state level.

The private banking sector. There are hundreds of commercial banks, of which the most important are the Deutsche Bank, the HypoVereinsbank, the Dresdner Bank, and the Commerzbank, though mergers have tended to shrink the number of major banks. Apart from conducting normal banking business, German banks provide financing for private businesses. As a result, the stock exchanges in Frankfurt, Düsseldorf, and other cities are less influential in providing finance for industry than parallel institutions in other countries.

Public and cooperative institutions. Germany has several types of public financial institutions, including credit and personal checking institutions and cooperative banks. Under public law, credit institutions operate as savings banks, and the state banks act as central banks and clearinghouses for the savings banks and focus on regional financing. The state-owned Kreditanstalt für Wiederaufbau ("Development Loan Corporation") channels public aid to developing countries.

The cooperative banks are headed by the DZ Bank (Deutsche Zentral-Genossenschaftsbank, or "German Central Cooperative Bank"), which serves as a central bank for some 1,500 industrial and agricultural credit cooperatives. There are also public and private mortgage banks, installment credit institutions, and the now-privatized postal check and postal savings systems, which were once operated by the federal postal service.

In East Germany the state bank was subordinate to the Ministry of Finance and designed to be a tool of central planning. It was part of a unified system that embraced not

only central and local government but also banks, insurance companies, and industries, all of which were directed in their use of funds.

With economic union on July 1, 1990, East Germany came under the central banking system of the Deutsche Bundesbank, which effected the conversion of the eastern system to the West German mark. Progressively, the western German commercial banks, insurance companies, and all the other financial institutions moved in. The ruined East German economy, the unemployment assistance fund, and the bankrupt state and local administrations all required massive financial transfusions from the federal government and the West German states. In stages, consumer subsidies have been removed, while wages, social insurance payments, and taxes have been progressively raised toward western levels.

TRADE

One of the world's leading exporters, Germany has consistently maintained a surplus with its trading partners. More than half of its trade is with members of the EU. Germany's principal export markets are France, the United States, the United Kingdom, Italy, and The Netherlands. Trade with eastern and central Europe has increased, and Germany has replaced the former Soviet Union and Russia as the primary trading partner for most countries in the region. Major exports include transport equipment (including automobiles), electrical machinery, and chemicals, as well as some food products and wine. Imports fall into remarkably similar categories, but in addition they include raw materials and semifinished products for industry. Germany's major sources of imports include France, The Netherlands, Italy, the United States, the United Kingdom, and Belgium.

Principal trading partners

Before unification East Germany specialized as a supplier of advanced industrial equipment, electronics, ships, and rail rolling stock to the communist bloc countries. Following economic unification, the countries of the former communist bloc were virtually unable to pay for equipment in hard currency, with disastrous consequences for eastern German industry. However, unlike the other former communist countries, eastern Germany, as part of united Germany, automatically received the benefits of full EC membership, though its factories also immediately faced overwhelming competition from western producers.

SERVICES

As is the case in many other countries with an advanced economy, Germany's service sector (*i.e.,* trade, transport, banking, finance, and administration) is a leading employer. This is abundantly clear in urban centres throughout western Germany, with their concentration of retailing, banking, and insurance. The transformation of eastern Germany along these lines is in progress, and the sector's importance has grown considerably there. For example, while the economies of most eastern and western German states were still dominated by manufacturing in the early 1990s, by the end of the decade a majority of states, and the country as a whole, had economies with a higher level of output by private firms providing services (even excepting trade and transport, which are categorized separately). In short, the German economy, for years one of the world's most manufacturing-oriented economies, has become dominated by services. This is particularly well illustrated by Berlin, where manufacturing's importance has declined sharply; indeed, the city has become an increasingly significant centre for both public and private international and national service-sector institutions.

Although foreign tourism to Germany is substantial, receipts from German tourists abroad exceed the receipts from foreign visitors to the country. In comparison with many of its neighbours, Germany does not rely heavily on tourism for income. The Alps and the Rhine and Moselle valleys are leading destinations, though urban areas (*e.g.,* Frankfurt, Munich, and Berlin) also attract many visitors, and local festivals in places such as Bayreuth also entice tourists. Tourism to eastern Germany, particularly to the beaches along the Baltic Sea, has increased significantly since unification.

LABOUR AND TAXATION

Germany's highly urban and industrialized character is reflected in its employment patterns. Services, including trade and finance, account for the largest share of employment. At the turn of the 21st century, about one-fifth of workers were employed in manufacturing, and fewer than 3 percent were employed in agriculture-related industries.

Prior to World War II most German labour unions were organized along partisan lines. After the war, however, trade unions were reconstituted to represent an entire industrial branch rather than simply a single trade or skill, thus avoiding interunion jostling within plants, and an independent German Trade Union Federation (Deutscher Gerwerkschaftsbund; DGB), which represents nearly all the country's unionized industrial employees, was established. The federation is an agglomeration of mostly blue-collar unions (though there are some white-collar unions), the largest of which are the United Service Industries Union (Vereinte Dienstleistungsgewerkschaft), the Metalworkers' Union (IG Metall), the Public Services and Transport Workers' Union (Gewerkschaft Nahrung-Genuss-Gastätten), the Mining, Chemical, and Energy Union (Industriewerkschaft Bergbau, Chemie, Energie), and the Federation of Civil Servants (DBB–Beamtenbund und Tarifunion).

Although Germany's social economy allows collective bargaining, unions are generally viewed as partners rather than opponents of business. The common interests of management and labour are expressed in works councils. Labour also has a right of codetermination (*Mitbestimmungsrecht*) through representation on managerial boards. About one-third of all German workers belong to a trade union. German's average labour costs are among the highest in the world.

Taxes are the major source of revenue for all levels of government. Five types of taxes—value-added, wage, assessed income, energy, and corporate—account for nearly four-fifths of all revenues. The federal government and the states each receive more than two-fifths of the principal taxes, leaving the remainder for local councils. A host of lesser taxes are specific to either the federal level (such as the tax on tobacco and alcohol and customs duties), the states (tax on beer and motor vehicle licenses), or the local authorities (tax on real estate, trade, and public entertainment). The states also benefit from property taxes. Because the taxing potential of the states is unevenly distributed, the economically weaker or smaller states share in the tax revenue of the richer or more populous states through a process of "horizontal financial equalization," which became an especially controversial matter after unification, when the poorer eastern German states became entitled to subsidies from western Germany. The federal corporate tax rate is about 25 percent, and, when local taxes are included, the overall tax burden reaches about 40 percent. Germany imposes a value-added tax of 16 percent to most goods and services. To spur economic growth, the German government reduced personal and business taxes in the late 1990s.

Types of taxes

The federal government is obligated to transmit certain revenues to the EU. Germany's disproportionately large payments to the EU have become a significant domestic and EU-wide political issue. As one of the world's richest countries, Germany feels obliged to supplement its regular contributions to the United Nations with complex international aid programs of its own.

TRANSPORTATION AND TELECOMMUNICATIONS

Germany has a dense network of communication facilities. Its geographic location in the heart of Europe also makes Germany responsible for facilitating the transit traffic serving neighbouring countries.

Waterways. The Rhine has the great advantage of having a remarkably even flow, with a spring-summer high water from the Alpine snowmelt supplemented by autumn-winter rains in the Central German Uplands. It is navigable from its mouth to above Basel, Switzerland, with the support in its upper course of the French Grand Canal d'Alsace. Typically, river transport is accomplished by

using push units propelling several barges. Since World War II the Rhine tributaries have been opened up for travel and transport. Navigation on the Moselle has been improved to the Saar region and Lorraine, on the Neckar to Stuttgart, and on the Main to provide a major European link to the Danube. Canals through the Ruhr region allow access to the northern German ports of Emden, Bremen, and Hamburg; waterway connections eastward to Berlin were once inadequate, especially at the crossing of the Elbe, but are being improved.

Seaports. Hamburg, which handles some one-third of the overall tonnage by weight, is Germany's principal port, accommodating the largest share of containers, as well as various ores and a wide range of general cargo. But because the largest tankers can no longer reach the Hamburg refining centre, Wilhelmshaven has become the prime destination for Germany's oil imports, as well as a major port in general. The Weser ports (Bremen and Bremerhaven) also handle a significant amount of total tonnage and containers; Bremen has an important general cargo trade. Although Hamburg, the Weser ports, and Emden are able to transship heavy goods to the interior by waterway, they play a less important role in this area than Rotterdam and other ports located at the mouth of the great Rhine waterway and closer to the Rhine-Ruhr area than the northern German ports are. Because the Elbe River leads to the port of Hamburg in what was West Germany and the Oder River to Szczecin (Stettin) in Poland, East Germany developed a new deep-sea port at Rostock, which was served by motorway and rail but had no waterway link. Some commodities needing fast service continued to arrive at special East German quays at Hamburg. Hamburg has regained much of its former Elbe trade since unification, but Rostock remains busy. Ferries for passengers, road vehicles, or railcars link Germany with Scandinavian destinations.

Railways. During the country's partition, the rail system was divided as well. In West Germany the Deutsche Bundesbahn (German Federal Railroad) reconstructed the old system, converting it to electric and diesel traction. The configuration of the country placed the emphasis on north-south routes. The burdened Rhine valley lines and the difficult routes through Hessen were augmented by a superbly engineered (and extremely expensive) high-speed track that permitted speeds up to 155 miles (250 kilometres) per hour. High-speed passenger rail service now links major German urban centres with one another and with other European destinations. The rail system competes successfully with airlines by offering fast and regular Inter City (IC) and Trans-European Express (TEE) trains.

East Germany retained the old name of Deutsche Reichsbahn ("German Imperial Railroad") for its system. Postwar reconstruction was slow, with efforts centring on rail links with the country's eastern European neighbours and the port of Rostock. The once-important east-west routes across the inner-German boundary were either removed or neglected. The Berlin outer-ring railroad was completed, enabling mainline and local traffic to avoid West Berlin. Unification revealed the dilapidated state of the system. Within Berlin, the trains, buses, and trams of the public transport were totally divided. Yet, when the border reopened, both the S-Bahn (Stadtbahn), an elevated railway system, and the U-Bahn (Untergrundbahn), the subway, were immediately able to resume service from east to west. (Two U-Bahn lines had continued to cross through areas of East Berlin but were not permitted to make stops at intermediate stations.) A lengthy and costly process of fully restoring a unified system, both within Berlin and nationally, began in late 1989 and resulted in significant progress for eastern Germany's railway network.

Highways. Germany completed the first section of the autobahn, near Berlin, in 1921, and several other countries quickly followed with their own versions of high-speed expressways. In the 1930s Hitler exploited the autobahn for economic, military, and propaganda purposes, but during World War II this German innovation—regarded as a model for modern expressways—was battered. The West German government greatly extended the system from 700 miles (1,125 kilometres) in 1950 to more than 5,000 miles (8,000 kilometres) by the time of unification. With powerful German automobiles able to cruise at their top speeds without speed limits, the autobahn gained an aura of automobile-centred romanticism throughout the world in the second half of the 20th century. However, road construction has encountered serious opposition from the country's environmentalist movement, and in inhabited areas the roads sometimes have been narrowed rather than widened to reduce traffic speed. Because the growth of the system has been slower than the growth of traffic, congestion is a serious problem, especially on motorways in industrial areas. Attempts to divert shipment of goods to the railways have not prevented a steady rise in the transport of goods by road. Western German motorways have direct trans-frontier connections with the similar systems of Denmark, The Netherlands, Belgium, France, and Austria.

The autobahn

With a lower growth rate of motor traffic (and an official policy of giving preference to the railroads), postwar construction of motorways was less advanced in East Germany. There were some improvements in central Germany, and new links to the ports of Rostock and Hamburg were constructed. The Berliner Ring, a circle of expressways around the city, was completed in 1979. With reunification, many transboundary roads were reopened and road surfaces improved. However, the construction of new roads has been hindered by conflicts between those seeking greater accessibility for automobiles and those seeking to protect the landscape and reduce air pollution.

Air transport. Germany's major long-distance airline is Lufthansa, though there also are a number of other carriers that service European and North American destinations. Frankfurt's airport, one of the world's busiest, is the country's largest; airports in Düsseldorf, Munich, and Berlin (Tegel) are also of major importance. During the period of partition, passenger traffic from West Germany to West Berlin was restricted to the airlines of France, the United Kingdom, and the United States. After unification Berlin was opened to German carriers (and indeed to carriers of other countries). East Germany discouraged internal air traffic and the growth of regional airports, using the rail and Berlin subway systems to serve its major international airport, Berlin-Schönefeld, south of the city. During the late 1990s, expansion of Schönefeld began, and it was expected to become united Berlin's only commercial airport by about 2010, after major expansion projects.

Telecommunications. After World War II West Germany developed an advanced telecommunications system. By contrast, the East German telephone system was completely insufficient; people requesting a telephone often were faced with a wait of up to 12 years. The deficiencies of the telecommunications system were a major impediment to the restructuring of the administration and the economy following unification, but by the late 1990s rapid reconstruction of the system using current technology made eastern Germany a world leader in advanced telecommunications infrastructure.

The leading German telecommunications company is Deutsche Telekom AG. During the late 1990s the entire sector was liberalized, increasing the number of telecommunications firms and competition for Deutsche Telekom from companies such as Vodafone and VIAG Interkom. The adoption of telecommunications services by German consumers has been widespread, particularly for cellular services. By the early 21st century more than one-third of the population used the Internet regularly.

Internet use

(T.H.El./G.H.K./W.H.Be.)

Administration and social conditions

The structure and authority of Germany's government is derived from the *Grundgesetz*, or Basic Law, which went into force on May 23, 1949, after formal consent to the establishment of the Federal Republic (known as West Germany) had been given by the military governments of the Western occupying powers and upon the assent of the parliaments of the *Länder* (states) to form the *Bund* (federation). West Germany then comprised 11 states and West Berlin, which was given the special status of a state without voting rights. The capital was located

in the small university town of Bonn, as an obviously provisional solution. Virtually simultaneously, on Oct. 7, 1949, the Soviet Zone of Occupation was transformed into a separate, nominally sovereign nation (if under Soviet hegemony), known formally as the German Democratic Republic (and popularly as East Germany). The five federal states (*Länder*) within the Soviet zone were abolished and reorganized into 15 administrative districts (*Bezirke*), of which the Soviet Sector of Berlin was the capital.

In the case of West Germany, full sovereignty was achieved only gradually: many powers and prerogatives, including those of direct intervention, were retained by the Western powers and devolved to the Federal Republic only as it was able to grow in economic and political stability and to be integrated into the Western community of nations. The tripartite offices of military governor were replaced upon the creation of the Federal Republic by those of high commissioners, and upon its achievement of full sovereignty on May 5, 1955, the high commissioners became ambassadors accredited to the president of the republic.

The Democratic Republic regarded its separation from the rest of Germany as complete, but the Federal Republic regarded East Germany as an illegally constituted state, until the doctrine of "two German states in one German nation" was developed in the 1970s. A sequence of gradual rapprochements between the two governments helped regularize the anomalous situation, especially concerning travel, transportation, and the status of West Berlin as an exclave of the Federal Republic.

The new *Länder*

During the process of unification, East Germany, as a condition for integration into the Federal Republic, reconstituted the five former historic states of Brandenburg, Mecklenburg-West Pomerania, Saxony, Saxony-Anhalt, and Thuringia. As states of the united Germany, they have adopted administrative, judicial, educational, and social structures parallel and analogous to those in the states of former West Germany. East and West Berlin were reunited, forming a state by itself.

Because the former East Germany de jure had petitioned to be integrated into the existing Federal Republic and de facto was thoroughly bankrupt and prostrate politically and economically, it entered the German Federal Republic without any bargaining power whatsoever. All changes and adaptations were thus completely on the terms set by Bonn, with few legal concessions allowed for the transitional period.

With the achievement of unification on Oct. 3, 1990, all remaining vestiges of the Federal Republic's qualified status as a sovereign nation were voided. No longer, for example, was Berlin still technically occupied territory, with the ultimate authority vested in the military governors. The very choice of the term *Grundgesetz*, or Basic Law, had purposely been chosen in 1949 to stop short of implying a permanent constitution (*Verfassung*), the promulgation and signing of which remains a final formal step in Germany's long progression to complete and unrestricted sovereignty.

In the days of the empire, German society was among the most intricately hierarchical in all Europe. The social upheaval of two major wars and economic change, though loosening this rigidity, left the basic class structure—with certain notable exceptions—essentially intact. Present-day German society is no longer plagued by class consciousness, but a sense of one's station in life is implicitly understood. Education still commands a greater awe in Germany than in many countries; the professor is held in an esteem incomprehensible to foreigners, and the title "doctor" is an all-but-essential credential for advancement not only in the upper echelons of the professions and the civil service, where a certain erudition is not inappropriate, but also—even more baffling to outsiders—in the ranks of business. The older authoritarianism, however, especially after the social upheaval of the 1960s and '70s, was greatly tempered by a healthy skepticism toward those who formerly simply by virtue of their title and position could command respect and obedience. As in other advanced nations, the basis of power has passed to the technical and managerial meritocracy.

By the time of unification in 1990, society and social conditions in western Germany had come in close harmony with those of its western European neighbours, while the inhabitants of former East Germany, after 40 years of one of the most rigid of Marxist-Leninist regimes, had grown radically apart from their kinsmen in the West. Collectivism and suppression of individual initiative in return for assured employment and provision of the basic requirements of life free or at low cost had produced a society antithetical to that in the West.

GOVERNMENT

The constitutional framework

The Basic Law has many affinities with the constitutions in the Anglo-American democracies and its predecessor, the Weimar Constitution (upon which it drew heavily). The parliamentary form of government incorporated many features of the British system, but, since West Germany, unlike Great Britain, was to be a federation, many political structures were drawn from the models of the United States and other federative governments. In reaction to the unitary state of the Nazi era, the Basic Law gave the states considerable autonomy, much of which has been eroded by constitutional amendments, fiscal developments, and a political insistence on uniform living conditions throughout the Federal Republic. In addition to federalism, the Basic Law has two other features similar to the Constitution of the United States: (1) its formal declaration of the principles of human rights and of bases for the government of the people and (2) the strongly independent position of the courts, especially in the right of the Federal Constitutional Court to declare a law unconstitutional and void.

Executive and legislative power. The formal chief of state is the president. Intended to be an elder statesman of stature, he is chosen for a term of five years by an assembly specially convened. His functions are far from being merely honorific. Apart from representing the Federal Republic among other nations and signing all federal legislation and treaties, he nominates the federal chancellor and the chancellor's Cabinet appointments, whom he may dismiss upon the chancellor's recommendation. He cannot, however, dismiss the federal chancellor or the Bundestag, the federal parliament. Among his other important functions are those of appointing federal judges and certain other officials and the right of pardon and reprieve.

The government in power is headed by the chancellor, who is elected by a majority vote of the Bundestag upon nomination by the president; in practice, the chancellor is always the chairman of his party. He is vested with considerable independent powers and initiates government policy. His Cabinet and its ministries also enjoy extensive autonomy and powers of initiative. The chancellor can be deposed only by an absolute majority of the Bundestag and only after a majority has been assured for the election of his successor. This "constructive vote of no confidence," as it is called, makes it unlikely for a chancellor or his government to be unseated, however much his working majority may have dwindled. Only twice has an attempt been made to unseat a chancellor, and only once with success (in 1982 when Helmut Schmidt was replaced by Helmut Kohl). The Cabinet may not be dismissed by a vote of no confidence by the Bundestag. The president may not dismiss a government or, in a crisis, call upon a political leader at his discretion to form a new government, the latter constitutional provision being based on the experience of the sequence of events whereby Adolf Hitler—against the better judgment of the Reich president of the Weimar Republic—became chancellor in 1933.

Ministries of the Cabinet

The number of Cabinet ministers may vary. Most Cabinet members are delegates to the Bundestag and are drawn from the majority party or proportionally from the parties forming a coalition, but the chancellor may appoint persons without party affiliation from a certain area of technical competence. These nondelegate members speak or answer questions during parliamentary debates.

The chancellor is immediately assisted by his secretaries of state, who administer various aspects of foreign and internal affairs or the conduct of press and information services; they may exercise wide powers of discretion in carrying out the instructions of the chancellor. The

Cabinet ministries—apart from the major areas of foreign policy, finance, defense, internal affairs, justice, and economy—are responsible for such technical and social functions as post and telecommunications (including the post office, the telephone system, and certain aspects of broadcasting); youth, family, and health; economic cooperation; education; nutrition; labour; and housing and urban development and regional planning.

Certain individual organs of government administer such areas as internal security, intelligence, press and public information, and statistics; they operate under the direct authority of the chancellor. The Federal Audit Office, independent of both chancellor and Bundestag, is charged with the accounting and budgetary control of all governmental functions.

Methods of constituting the parliamentary chambers

The Bundestag is the cornerstone of the German system of government; as of the general election of Dec. 2, 1990, it consisted of 662 members (subject to slight variation). Its delegates are chosen either in general elections held every four years or in special by-elections. In addition to the members elected by each district, a set of state delegates at large is elected simultaneously, both as a means of ensuring stability and continuity of representation by the major parties in the lower chamber and as a corollary to look to the interests of nation, state, party, or bloc of voters at large. In this latter function the delegates at large serve as a counterbalance to the parochial tendencies inherent in strict representation by district constituency mandate.

The Bundestag exercises much wider powers than does the upper chamber, known as the Bundesrat, or Federal Council. In the Bundesrat the states themselves exercise authority to protect their rights and prerogatives. Its members are appointed by the governments of the states, each state sending from three to five members, depending on size and population; the new state of Berlin, for example, sends four members. The delegations are bound by the instructions of their provincial governments. All legislation originates in the Bundestag, and the consent of the Bundesrat is necessary only on certain matters directly affecting the interests of the states, especially in the area of finance and administration and for legislation in which questions of the Basic Law are involved. It may exercise a restraint on the Bundestag by rejecting certain routine legislation passed by the lower chamber, but, unless the bills fall within certain categories, its vote may be overridden by a simple majority in the Bundestag. Should the president be absent abroad for extended periods or withdraw from office, the speaker of the Bundesrat may deputize for him.

The powers of the Bundestag are kept in careful balance with those of the Landtage, the state parliaments. Certain powers are specifically reserved to the republic—foreign affairs, defense, currency and minting, post and telecommunications, customs and problems of international trade, and matters affecting citizenship. The Bundestag and the states may pass concurrent legislation in such matters when it is necessary and desirable, or the Bundestag may set out certain guidelines for legislation; drawing from these, each individual Landtag passes appropriate legislation in keeping with the particular needs and circumstances of its own state. In principle, the Bundestag initiates or approves legislation in matters in which uniformity is essential, but the Landtage otherwise are free to act in areas in which they are not expressly restrained by the Basic Law.

Provincial and municipal government. Certain functions are expressly the province of the states, notably education and law enforcement; yet even here an attempt is made to maintain a degree of uniformity among the 16 states through joint consultative bodies. The governments of the states are generally parallel in structure to that of the Bund but need not be. In 13 of the states the head of government has his own Cabinet and ministers; each has its own parliamentary body, but in the city-states of Hamburg, Bremen, and Berlin the mayor is simultaneously the head of government of the state. The municipal senates serve also as provincial parliaments, and the municipal offices assume the nature of provincial ministries.

The administrative subdivisions of the states (exclusive of the city-states and the Saarland) are the *Regierungsbezirke* (administrative districts). Below these are the divisions known as *Kreise* (counties) or, in parts of northern Germany, as *Grafschaften,* roughly equivalent to counties. Larger communities enjoy the status of what in Great Britain was formerly the county borough. The counties themselves are further subdivided into the *Gemeinden,* roughly "communities" or "parishes," which through long German tradition have considerable autonomy and responsibility in the administration of schools, hospitals, housing and construction, social welfare, public services and utilities, and cultural amenities.

JUSTICE

Structure of the court system

State and federal courts. The German court system differs from that of some other federations, such as the United States, in that all the trial and appellate courts are state courts, while the courts of last resort are federal. All courts may hear cases based on law enacted on the federal level, though there are some areas of law over which the states have exclusive control. The federal courts assure the uniform application of national law by the state courts. In addition to the courts of general jurisdiction for civil and criminal cases, the highest of which is the Federal Court of Justice, there are four court systems with specialized jurisdiction in administrative, labour, social security, and tax matters. The jurisdiction of the three-level system of administrative courts extends, for example, to all civil law litigation of a nonconstitutional nature unless other specialized courts have jurisdiction.

While all courts have the power and the obligation to review the constitutionality of government action and legislation within their jurisdiction, only the Federal Constitutional Court in Karlsruhe may declare legislation unconstitutional. Other courts must suspend proceedings if they find a statute unconstitutional and must submit the question of constitutionality to the Federal Constitutional Court for its determination. In serious criminal cases the trial courts sit with lay judges, similar to jurors, who are chosen by lot from a predetermined list. The lay judges decide all questions of guilt and punishment jointly with the professional judges. Lay judges participate also in some noncriminal matters.

Judges play a more prominent and active role in all stages of legal proceedings than do their common-law counterparts, and proceedings in German courts tend to be less controlled by prosecutors and defense attorneys. There is less emphasis on formal rules of evidence, which in the common-law countries is largely a by-product of the jury system, and more stress on letting the facts speak for what they may be worth in the individual case. Observers from common-law countries might find the lack of certain elements of the adversarial process, particularly in criminal trials, difficult to reconcile with their notion of fairness, but Germans are not necessarily persuaded that such a process leads to greater justice. There is no plea bargaining in criminal cases. In Germany, as in most European countries, litigation costs are relatively low as compared with the United States, but the losing party in any case usually must pay the court costs and attorney fees of both parties.

Administrative law

While codes and statutes are viewed as the primary source of law in Germany, the role of precedent is of great importance in the interpretation of legal rules. German administrative law, for example, is case law in the same sense that there exists no codification of the principles relied upon in the process of reviewing administrative action. These principles are mostly judge-made law. Germans see their system of judicial review of administrative actions as implementation of the rule of law. In this context, the emphasis is on the availability of judicial remedies. While in the United States there is a tendency to consider procedural safeguards, such as a requirement for a hearing before an administrative decision is made, to be a major requirement of the rule of law, a citizen's rights in this formative stage have been of lesser concern in Germany.

Integration of former East Germany. The integration and adaptation of the administration of justice of the new

states into the system operative in the Federal Republic was greatly complicated by an unavoidable incapacitation of a large number of judges. They had to be dismissed either because they owed their appointment as judges mainly to their loyalty to the old SED or because of their records. To fill the many vacancies created in the courts of the new *Länder*, judges and judicial administrators had to be recruited from former West Germany. A large number were "put on loan" from the western *Länder* and many others urged out of retirement to help during the transition.

POLITICAL INSTITUTIONS

The electorate. Both the quadrennial general and provincial elections as well as local elections are attended with the greatest interest and involvement on the part of the electorate. The public is kept informed on political issues through intense coverage in the press, television, and radio, and political affairs frequently provide a topic for debate among German citizens. In 1970 the voting age was reduced from 21 to 18. Although voting is not compulsory, an extremely high percentage of citizens participate. Since elections in the states are staggered throughout the life of each Bundestag, they act as weather vanes of public opinion for the incumbent federal government.

Voting age

Political parties. Germany's political parties, the sheer proliferation of which contributed to the downfall of the Weimar Republic in 1933, tended to consolidate during the early days of the Federal Republic. Smaller parties either allied themselves to the larger ones, shrank into insignificance, or simply vanished. Reunified Germany has, in effect, only two numerically major parties, the Christian Democratic Union (CDU) and the Social Democratic Party of Germany (SPD), neither of which can easily attain a parliamentary majority. In addition, there are three numerically small but powerful parties, the Christian Social Union (CSU, the Bavarian sister party of the CDU), Alliance 90/the Greens (the Greens), and the Free Democratic Party (FDP). Two minor parties, the Party of Democratic Socialism (PDS) and the German People's Union (DVU), hover above or below the threshold of representation in the Bundestag or the state diets set by the "5 percent rule," whereby a party may send delegates only if it wins at least 5 percent of the vote in a given election. This rule has proved a highly effective instrument in excluding radical parties of whatever stripe and in preventing the formation of splinter parties. Since 1961 the major parties have had to hold their power through coalitions with an additional party. Dissent within the major parties is contained in the wings and factions of each respective party.

The Christian Democratic Union (Christlich-Demokratische Union) is the party of the centre-right—the "bourgeois" party in older European terms. It headed Bonn governments from 1949 to 1966 and governed from 1982 to 1998 in alliance with the FDP and CSU. It is established in all states except Bavaria, where the more conservative CSU (Christlich-Soziale Union) functions as its counterpart in effectively a permanent coalition.

Party constituencies

In its origins, the Christian Democratic Union represents a merger of the old Catholic Centre Party with kindred bourgeois parties, either Protestant or nonsectarian. In a nation in which one's religion had also often been one's politics, the party's strongest constituencies are still in the Roman Catholic districts, although the sectarian Christian aspect is of only incidental emphasis, chiefly among older voters. The party's policies emphasize a free-market economy, national unity, and Germany's place in the Western community, the European Union (EU), and the North Atlantic Treaty Organization (NATO).

Throughout the existence of the Democratic Republic, an eastern branch of the CDU (known as CDU-Ost) had been tolerated to preserve the facade of a multiparty system. With the overthrow of the ruling SED in East Germany's first free elections on March 18, 1990, it was this rump party that took power by a large mandate, with Lothar de Mazière as minister president presiding over the six-month transitional period to unification.

The Social Democratic Party of Germany (Sozialdemokratische Partei Deutschlands), which governed Germany in coalition with the FDP from 1969 to 1982 and regained power in coalition with the Greens in 1998, is the heir to the Marxist parties dating from the 19th century. In East Germany, which historically had been a stronghold of the socialist movement in Germany, the SPD had been subsumed by the Socialist Unity Party (Sozialistische Einheitspartei Deutschlands, or SED), in 1946. Ostensibly a combination of the old Communist Party of Germany (Kommunistische Partei Deutschlands) and the Socialist Party, the SED was in fact simply the ruling communist party. In West Germany the SPD's early postwar leadership, drawing strength from its record of opposition to Nazism, maintained a rigorous policy of adherence to classic Marxist doctrine, vehement opposition to the communist movement from which it had split in the early 20th century, and rejection of rearmament for West Germany and its integration into the Western military defense system. In 1959, however, the SPD, in the Bad Godesberg Resolution, changed its approach: the call for nationalizing large industries was forsaken in favour of gradualist reform, and appeals to class warfare were abandoned. The party was able, thereby, to attract greater middle-class support, whether for reasons of ideology or as an alternative to continued CDU leadership.

In its approach to unification, the SPD called for caution and has profited from a backlash against the high financial and emotional costs of the rush to unification spearheaded by the CDU-CSU. While the SPD suffered electoral losses immediately after reunification, it replaced the CDU-CSU as the leading party in the 1998 federal elections.

Because neither the CDU-CSU nor the SPD could generally elect enough delegates to form a viable majority in the Bundestag, the balance of power, with short exceptions, was in the hands of the Free Democratic Party (Freie Demokratische Partei) through coalitions with either of the larger parties. The FDP is the successor of older German liberal parties with a free-trade, pro-business, and anticlerical cast. Today it serves as a liberal, bourgeois alternative to the CDU and SPD and sometimes exercises a power beyond the 5 to 10 percent it commands from the electorate. A small counterpart party that existed in the East, the Liberal Democratic Party (Liberal-Demokratische Partei) lent surprising strength to the FDP in the 1990 election.

The Greens

Alliance 90/the Greens (Bündnis 90/die Grünen), an environmentalist party formed in 1993 through a merger of the East German Alliance 90 (founded 1991) and the West German party the Greens (founded 1979), is the only successful completely new party of the postwar era. Formed by mainly younger groups of environmentalists, opponents of nuclear power, feminists, and pacifists, the Greens successfully broke the 5 percent barrier in 1983 by winning in by-elections. The Greens advocate expanded rights for minorities as well as environmentally conscious policies, and they tend to appeal to young, educated voters. After losing support in 1990, the Greens benefited from growing disillusionment with the CDU-CSU government during the mid-1990s. In 1994 the Greens replaced the FDP as the third most popular party, and in 1998 they displaced the FDP from its traditional political role by forming a governing coalition with the SPD.

The Party of Democratic Socialism (Partei des Demokratischen Sozialismus, or PDS) is the successor party to the former Socialist Unity Party of Germany. It advocates state intervention in the economy on behalf of the disadvantaged and expresses the grievances of easterners who have not benefited from unification. It exceeded the 5 percent barrier in the 1998 federal elections and commands about 20 percent of the vote in eastern Germany.

Of the small and fringe parties, only the rightist-nationalist groups the Republican Party (Republikaner) and the German People's Union (Deutsche Volksunion), together with a handful of special-interest bodies, are visible in national or regional elections, but they have been unable to surmount the 5 percent barrier in most states or in the Bundestag. In an exception to the ban, a small party of the Danish minority, the Southern Schleswig Voters Association, sends one delegate to the Schleswig-Holstein state diet.

ARMED FORCES

The Federal Republic has been a member of the North Atlantic Treaty Organization (NATO) since the implementation of the Paris Treaties in May 1955. Until unification West Germany was the only NATO country of western Europe with territory bordering two member nations of the Warsaw Pact, and planning from the early days of the Western defense system was posited on West Germany's vulnerability to an armed invasion and to becoming the possible site of a land war.

The German contribution to the Western defense system, apart from playing host and contributing to the continued presence of Allied troops on its soil, takes the form of its combined arm of defense known as the Federal Armed Forces (Bundeswehr). Constituting the largest contingent of NATO troops in Europe, the armed forces are divided into an army, navy, and air force. From its inception it was envisioned as a "citizens' " defense force, decisively under civilian control through the Bundestag, and its officers and soldiers trained to be mindful of the role of the military in a democracy. Conscription for males is universal, the military liability beginning at 18 and ending at 45 years of age. The period of compulsory active service is 18 months. The right to refuse to enter military service on grounds of conscience is guaranteed under Article 4 of the Basic Law; instead, a recognized conscientious objector must perform 20 months of socially useful service. After unification the exercise of this right increased nearly twofold.

The Federal Armed Forces

The Federal Republic maintains a separate Coast Guard and Border Patrol. As a concession to Bavaria's once-special position within the old German Empire, this force, although maintained by the federal government, is still known there as the Bavarian Border Patrol.

Upon unification the former People's Army (Volksarmee) of the Democratic Republic was integrated into the Federal Armed Forces, with officers from the west assuming top command and overseeing the assimilation into a NATO force of what had been implacable enemies. The special Border Troops set up to guard the Berlin Wall and boundary with former West Germany, together with the factory militia, were disarmed and dissolved.

LAW ENFORCEMENT

There is no nationwide German police force, and law enforcement remains a province reserved to the states. Each state maintains its own force, which is charged with all phases of enforcement, except where its function is assumed by a municipal police force. In a state of national emergency the federal government may commandeer the services of various state police units, together with the standby police reserve that is trained and equipped by each state for action during civil emergencies. Federal officers investigate certain actions, however, notably those inimical to the security of the state or criminal actions that transcend the confines of any given state. The Federal Crime Agency (Bundeskriminalamt) assists the federal and state units as a clearing agency regarding criminals and criminal actions.

The former People's Police of the Democratic Republic was dissolved upon unification, and its members integrated into the police force of the new states. The loathed Ministry for State Security (Ministerium für Staatssicherheit, popularly known as Stasi), the offices of which had been stormed in popular uprisings and whose files had been removed into Western custody, was dissolved.

EDUCATION

Preschool, elementary, and secondary. Schooling is free and compulsory for children 6 to 18 years of age. While control of education is the sovereign prerogative of the states, a permanent commission strives for a certain uniformity in curriculum, requirements, and standards, the implementation of which may vary, depending on the priorities of the individual state. Some of the books and study materials are free, and financial assistance and other forms of support are available in cases of hardship.

Preschooling, to which the notably German contribution in modern times is enshrined in the universal word kindergarten, can begin at 3 years of age. All children attend the *Grundschule* ("basic school") from age 6 until about age 10. Somewhat less than half continue elementary schooling in a junior secondary school called the *Hauptschule* ("main school") until about age 15 or 16, at which time they are assigned to a *Berufsschule* ("vocational school"), attended part-time in conjunction with an apprenticeship or other on-the-job training. This program makes it possible for virtually every young person in the vocational stream to learn a useful skill or trade, constantly adapted to the actual demands of the employment market.

Children who receive a commercial or clerical education, somewhat less than one-third of the school-age population, attend an intermediate school called the *Realschule* (roughly meaning practical school) and earn an intermediate-level certificate that entitles them to enter a *Fachschule* ("technical" or "special-training school"), completion of which is a prerequisite for careers in the middle levels of business, administration, and the civil service.

Approximately one-fourth of all children are chosen for study at the *Gymnasium* (senior secondary school, equivalent to a grammar school in Great Britain), in which a rigorous program lasting for nine years (levels 5 to 13) prepares them—with emphasis variously on the classics, modern languages, mathematics, and natural science—for the *Abitur* or *Reifezeugnis* ("certificate of maturity"), the prerequisite for matriculation at a German university. The traditional structure of the German *Gymnasium* has mainly shifted from being built around a single branch of studies to offering a "reformed upper phase" with a choice of courses.

Gymnasium

A small number of so-called *Gesamtschulen* (equivalent to British comprehensive schools) are now operated in each state. These are intended as an alternative to the previously rigid division into three levels, often criticized for forcing the choice of a child's future at too early an age, a choice that, once entered upon, was almost impossible to change. These schools offer a large range of choices and permit pupils more freedom in seeking the level best suited for them.

Higher education. The German universities, famed in history and noted for their enormous contributions to learning, especially in the 19th and early 20th centuries, have been put under severe strain by the swelling numbers of students and changing social conditions that have taxed the traditional structures of the universities beyond their capacities or accustomed functions. Today it has become all but impossible for students to take as long as they wish to complete their studies or to move from university to university as they please. Lecture rooms, seminars, and libraries are disastrously overburdened, and the explosion of knowledge and of higher education has undermined the usefulness of original research.

To meet the rapidly rising demand for higher education, the Federal Republic has steadily added to the number of existing universities. It has created entirely new academic universities to add to the ranks of the ancient universities, and it has upgraded the status of institutes and colleges of technology, education, and art to university rank, while creating new specialized or technical institutions such as the *Fachhochschule,* a higher technical college specializing in a single discipline, such as engineering, architecture, design, art, agriculture, or business administration. Little or no difference in prestige is attached to whether a student has studied at Heidelberg, founded in 1386, or at the new multimedia university at Hagen, Westphalia, established in 1976, where the teaching is accomplished largely by correspondence through regional study centres. In 1991 Germany had 97 universities and 200 institutions of equivalent rank. The doctorate is the only degree offered as such (although the Magister, or Master of Arts, abandoned in the 17th century, has been partially revived). The rough equivalent of a bachelor's degree is a *Diplom* or completion of a *Staatsexamen* (state examination).

There is an extensive range of possibilities for extended education or extramural studies. About 860 *Volkshochschulen* (adult education centres) enroll some 5.2 million adults for complete courses or individual subjects, whether in preparation for, or furtherance of, a career or out of personal interest. The government has also pro-

Volkshochschulen

moted the retraining and further vocational education of workers.

Problems of transition. Integration of the educational system of the former Democratic Republic brought a host of problems. Since the focus of education was to inculcate the values of the communist state, even the textbooks and some of the school materials were unsuitable for the educational aims of the Federal Republic. English replaced Russian as the first foreign language, rendering untold numbers of Russian teachers useless and creating a shortage of qualified English teachers. The reorientation of primary and secondary school teachers to the standards and aims of the republic became an important concern.

At the university level the issue of competence became acute. Since many of the faculty of East Germany's universities had been appointed for their soundness in Marxism-Leninism or loyalty to the SED, their qualifications became obsolete upon unification. The federal authority for qualifying universities to confer degrees and diplomas was hard put to give recognition to some institutions of university rank, while the ministers of education of the new states were subjected to great pressures to reconfirm existing appointments. Whereas in former East Germany research institutions had generally been separate from universities, the Western system combining research and teaching was implemented nationally following unification.

HEALTH AND WELFARE

Germany's system of social benefits is one of the most elaborate and all-embracing in the world. The country's pioneer work in social legislation was initiated in the 1880s to cover health and accident insurance, workers' and employees' benefits and pensions, miners' insurance, and the like. It has served as a model for similar programs in other countries.

Compulsory national and private insurance plans

Insurance and services. Health and retirement insurance are compulsory for all workers and employees earning below a certain level of income geared to the cost of living. Under German labour law, a categorical distinction is made between hourly wage earners and salaried employees, and differing rules and rates apply to each group. Employees above a certain salary level or self-employed persons are generally exempt from most obligatory payment systems; however, although the former usually participate in a firm's retirement plan, almost all persons in the upper salary brackets or the self-employed are covered by private insurance as comprehensive as the government-sponsored plans. With increasing prosperity, an increasing percentage of the population is subscribing to these somewhat more expensive but also more generous private plans. Nearly 90 percent of the population is covered by compulsory health insurance, and the Federal Republic ranks highest among comparable nations of continental western Europe in the proportion of money allowed toward health costs—about 90 percent of all medical costs incurred. Contributions range from about 8 to 12.5 percent of wages or salaries.

Medical care

Medical care is of a high standard, and rural areas are well served. Hospitals are usually operated by municipalities or religious bodies or as proprietary institutions owned by one or more physicians. Public health standards are high. A triumph of the public health system has been the conquering of tuberculosis, a disease formerly endemic to Germany but now rarely encountered. The extensive health-care system operative in the former Democratic Republic, in which universal free health care, medication, child care, nursing care, and pensions were funded by an obligatory state insurance system, underwent a reorganization from the exclusive management of the system by the trade unions to an alignment with the various employment, health, and retirement insurance systems of the Federal Republic.

Accident and retirement insurance are tied to health- and medical-care plans. The rate of compulsory accident insurance rises with the risks involved in one's job. The three major pension plans cover miners (the oldest, dating from Bismarck's introductory social legislation), workers, and employees. In general, men are eligible for retirement at 65 and women at 60, but both sexes may retire earlier or later under special circumstances.

Additional benefits. The Federal Republic provides several special systems of coverage for such special groups as war widows, orphans, and farmers. Unemployment insurance is funded through deductions from wages and salaries. Allowances are made for families with one or more children. Additional public allowances are granted to persons suffering disabilities from wartime injury, whether as military personnel or as civilians. Some small indemnification has been made to property owners whose holdings lay in former German territories now outside the Federal Republic. (For a discussion of property rights with regard to German reunification, see above *The economy: Problems of economic unification.*)

War reparations. Under agreements concluded with 12 European nations, the Federal Republic has paid compensation to the nationals of those countries who were victims of Nazi oppression or to their families and successors. The Federal Republic from the time of its foundation assumed the immense financial responsibility of making restitution to the Jewish victims of National Socialism. Claims for property confiscated during the Third Reich have been honoured, and Jewish refugees and expellees from that era, the vast majority of whom reside abroad, have been paid indemnifications and pensions. Massive reparations have been paid to Israel in the name of the Jewish people at large. The former Democratic Republic, on the other hand, ignored all such claims until 1990, when the transition government of Lothar de Mazière undertook similar restitution.

HOUSING

Private and government roles. Most of the capital for new housing is met by the private sector. Assistance is provided by the federal government to all citizens wishing to avail themselves of a building savings policy. After a certain minimum contribution by the policyholder has been deposited over a set period of years, the government provides a housing loan on generous terms for those wishing to build their own house or buy an apartment.

"Social housing"

Much of the housing built with government subsidies is devoted to "social housing," dwellings provided at "cost rent" far below the market rental value for families with many children, the disabled, the elderly, and persons with low incomes. More stringent definitions of tenants' rights, including injunctions against arbitrary or unfair evictions and protection against precipitous rent increases, seek to balance the rights of tenants with those of landlords.

The rebuilding of the cities in the 1950s and '60s, coupled with increased automobile ownership, invariably led to an initial emptying of the older city centres. Moves to counter the trend away from the inner city have been made, however, in the form of easier access and parking near the town centres, as well as improved public transport, large-scale refurbishing of ancient buildings, and the creation of pedestrian zones, with special entertainments, festivals, and attractions to lure the public back to town in the evening.

The physical appearance of villages and towns throughout West Germany was improved on a grand scale during the 1970s and '80s through extensive renovation programs undertaken by the states; grants, subsidies, and matching funds were made available to restore the exteriors of historic monuments and older buildings to their pristine condition.

The former eastern territories. In principle, under the communist government of the Democratic Republic each citizen and his or her family had the right to adequate accommodations. Rents everywhere, together with charges for heating and electricity, were held at extremely low, subsidized levels. The need for new housing was solved by erecting massive apartment blocks of cheap material.

Upon unification, the decrepitude of most older buildings in the east came as a shock to visitors from the west. With the exception of a few showpieces, the great majority of urban housing constructed between 1871 and 1914 (see above) seemed to have gone unpainted or unrefurbished since before World War II. The contrast of the gray, cheerless, and shabby cities, towns, and villages of the former Democratic Republic with the picture-postcard

lustre of the western sector was an eloquent testimony to the divergences that had to be overcome.

STANDARDS OF LIVING

The standard of living in the old states of the Federal Republic ranks as one of the highest in the world. The distribution of wealth compares favourably with that of other advanced nations. Powerful incentives to save are offered by the state not only in the form of housing subsidies and tax concessions but also through bonus saving schemes. Thus, for persons whose income does not exceed a certain level, savings of up to a fixed amount kept in a bank or savings institution for six or seven years will be granted a bonus of 14 percent of the principal by the government. Under the "DM 624 Law" the accumulation of capital assets is encouraged under a plan whereby workers below a certain earning level who agree to pay 52 deutsche marks per month (or 624 deutsche marks per annum) into a longer-term savings agreement, such as a home-savings contract, are given a "worker savings grant" by the state. Altogether 20 million workers participate in the federal asset-building program and a further 12 million in the state savings grant. The government has contributed some 140 billion deutsche marks to the asset-building scheme, through which savings of 600 billion deutsche marks have been accumulated by participants.

Savings incentives

Earning power for both workers and employees assures an adequate income to meet the cost of living. There is no exaggerated difference between the earning power of blue-collar workers and white-collar employees, salaries tending to be only moderately higher than wages. While above a certain level of management incomes and benefits rise precipitously, chief executive officers in Germany earn only 6 to 7 times the average worker's pay (as opposed to the vastly higher ratio in the United States). Since, as in all the EC countries, a major portion of tax revenue derives from excise levies and the value-added tax, low- and medium-income workers collectively bear the greater tax burden.

The absorption of the East German population and economy into that of the Federal Republic had no more than a marginal effect on living standards in the regions of the western sector despite a rise in unemployment and a housing shortage. Even the exorbitant cost of unification that became apparent only after the borders disappeared (and brought about a much-resented tax increase) seemed to cause few changes in western Germany. The deutsche mark held its strength and grew even stronger. By contrast, the introduction of the West German currency in East Germany in July 1990, far from being the "magic bullet" hoped for, tended to have a depressing effect at the microeconomic level. The eastern population with its much lower earning power suddenly had to pay Western prices for food and other commodities. The wholesale shutdown of former state factories and enterprises caused vast unemployment in especially the industrial cities of Thuringia, Saxony-Anhalt, and Saxony and resulted in much hardship and discontent.

Effects of unification on the standard of living

The 40 years of the Democratic Republic had produced a society in which employment was guaranteed and in which most of the requirements of life were provided free or at low cost. The demands of work were slight; competitiveness, initiative, and individuality were not qualities highly prized in a Marxist-Leninist society. The working population was thus ill-suited—even more than, say, the people of Poland, Czechoslovakia, or Hungary—for a blind, cold plunge into a free-market economy. But, unlike those in the latter countries, the East Germans had a rich and strong next-door relation that was sure to bail them out.

Cultural life

THE CULTURAL MILIEU

During the partition, the Federal Republic, as heir to Germany's older regions, was custodian to the greater portion of the country's rich cultural legacy. The major wealth of Germany's architectural monuments—of Roman Germany, of medieval Romanesque, of south German Baroque—fell within the borders of West Germany after World War II, as did many of the great libraries, archives, and facilities for the performing arts. Yet some of the greatest monuments of Germany's cultural and historical achievement were located in the German Democratic Republic, including the Wartburg of Luther near Eisenach, the Weimar of Goethe, the Leipzig of Bach; a large share of prewar Germany's art treasures rested in East Germany, especially in East Berlin and Dresden. After the division of Germany, many of the cultural assets originally from the eastern sector were removed to the west. Many of East Germany's artists, writers, and institutions, including entire publishing houses, transplanted themselves to West Germany or set up successor organizations there.

Yet, despite the political division, the German cultural and artistic tradition remained identifiably the same. In the German-speaking world a writer or painter or composer or playwright or sculptor was still German, whether holding a passport from the Federal Republic or from the Democratic Republic. Moreover, in the matter of art and literature, the adjective *deutsch* ("German") still knows no strict political boundaries. The Austrian Gustav Mahler, for example, is called a "German" composer and the Swiss Friedrich Dürrenmatt is a "German" playwright because they are in the German cultural tradition.

During the 40 years of separation it was inevitable that some divergence would occur in the cultural life of the two parts of the severed nation. Both West Germany and East Germany followed along traditional paths of the common German culture, but West Germany, being obviously more susceptible to influences from western Europe and North America, became more cosmopolitan. Conversely, East Germany, while remaining surprisingly conservative in its adherence to some aspects of the received tradition, was powerfully molded by the dictates of a socialist ideology of predominantly Soviet inspiration. Guidance in the required direction was provided by exhortation through a range of associations and by some degree of censorship; the state, as virtually the sole market for artistic products, inevitably had the last word.

Continuity and divergence in East and West

Admirably enough, the cultural commissars of the Democratic Republic steadfastly upheld certain cultural monuments that fell within East German borders—even though their provenance was regal, aristocratic, liberal, bourgeois, or religious and the content hardly reconcilable with the aspirations of the "State of Workers and Peasants": the Goethe House and Goethe National Museum on the Frauenplanstrasse in Weimar were carefully restored after the war and meticulously maintained; the Thomanerchor at the St. Thomas Church in Leipzig, the boys' choir made famous by Johann Sebastian Bach, continued to perform the cantatas and motets of the master in exactly the style of two and a half centuries previously; Dresden, although devastated by wartime bombing, made it an early priority to restore its opera house; and the music ensembles of the Democratic Republic, especially the Dresden Philharmonic and the Dresdner Staatskapelle, together with the Gewandhaus and Rundfunk orchestras of Leipzig, actually remained part of the mainstream of European music, going on tour in the West and freely exchanging performers, conductors, and producers.

It has been remarked that, during their separation, the two Germanys diverged not at all in music and only slightly in literature and the theatre but sharply in architecture and the plastic arts.

STATE OF THE ARTS

Government and audience support. For four centuries Germany has enjoyed a tradition of governmental support of the arts. Before the founding of the German Empire in 1871, the many small kingdoms, principalities, duchies, bishoprics, and free cities that preceded it—as well as Austria and German-speaking Switzerland—supported the arts; they established theatres, museums, and libraries, and their leaders acted as patrons to poets, writers, painters, and performers. The institutions thus founded and the convention of generous public support have continued uninterrupted to the present.

The quantitative dimensions of Germany's cultural life astound foreigners. In the Federal Republic a few hundred theatres are subsidized by the federal government, the

Theatrical and musical traditions

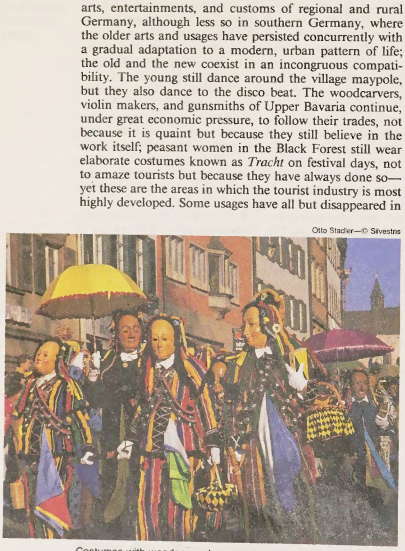
states, and the cities, in addition to the many privately financed theatres. Unlike the United States, Britain, and France, in which theatre is more often than not centred in one city, no one city in Germany dominates over the others. Also, productions in Vienna and in Zürich, Switz., are significant to the artistic life of the Federal Republic, and artists and resources move easily and freely among the theatrical and operatic companies within the German-speaking regions. Only in Vienna, the capital in which the arts arouse far more intense passions than do politics, does theatre have a broader audience base than in Germany. Audiences in Germany are not limited to a small intellectual or social elite but are drawn from all ranks of society. Season tickets, group arrangements, bloc tickets bought by business firms, and theatre clubs constitute the major patronage of such production companies as the People's Independent Theatre (Theater der Freien Volksbühne), dating from 1890 in Berlin. Going to the theatre or opera in Germany is about as affordable and as unremarkable as attending the cinema is elsewhere.

The same is true of concert music. Every major city has one or more symphony orchestras offering many concerts and recitals each week; in the smaller cities and towns the music and concert fare is less well provided for only in terms of quantity and, perhaps, professional quality.

In few countries of the world are the arts so lavishly cultivated as in the Federal Republic in terms of the proliferation of cultural amenities, the funds allotted to them, and the attendance upon them. While this abundance and generous support has not called forth a new era of brilliance able to rival that of the Weimar Republic—when Germany (especially Berlin) experienced a resurgence in the arts and a proliferation of creative talents unparalleled since German Classicism and Romanticism—there are a number of noteworthy individual talents and movements dotting the contemporary landscape.

Traditional arts and crafts. The incursions of modern patterns of life have done much to weaken the traditional arts, entertainments, and customs of regional and rural Germany, although less so in southern Germany, where the older arts and usages have persisted concurrently with a gradual adaptation to a modern, urban pattern of life; the old and the new coexist in an incongruous compatibility. The young still dance around the village maypole, but they also dance to the disco beat. The woodcarvers, violin makers, and gunsmiths of Upper Bavaria continue, under great economic pressure, to follow their trades, not because it is quaint but because they still believe in the work itself; peasant women in the Black Forest still wear elaborate costumes known as *Tracht* on festival days, not to amaze tourists but because they have always done so—yet these are the areas in which the tourist industry is most highly developed. Some usages have all but disappeared in the villages: older women now seldom wear black dresses and scarves, and the village men no longer appear in top hat and cutaway for a funeral procession.

Otto Stadler—© Silvestris



Costumes with wooden masks worn during the pre-Lenten celebrations in Rottweil, Baden-Württemberg.

Popular festivals

Popular festivals still abound in the west, southwest, and south, the regions that have clung most to the practices of a traditional, preindustrial age. Near-heathen usages such as the donning of elaborate wooden masks during the pre-Lenten celebrations in the southwest remain unaffected in spite of being televised; hundreds of smaller towns and larger villages in the south still commemorate an anniversary from the Thirty Years' War by a parade in 17th-century costume or, in Roman Catholic areas, march in full procession on Corpus Christi Day. What is remarkable is not merely that these traditions survive but also that the homelier and less celebrated of them remain truly genuine and naive in the observance.

Literature and theatre. German literature holds less than its deserved status in world literature in part because the lyrical qualities of its poetry and the nuances of its prose defy the most inspired translation. Even the most sublime of figures in German literary history, such as Goethe and Schiller, are doomed to remain known to the world outside the German regions largely by reputation. In the 20th century perhaps four German poets and writers have won a permanent niche in world literature—Franz Kafka, Thomas Mann, Rainer Maria Rilke, and Bertolt Brecht, all of whose works date from the early decades. Two German novelists have won a popular following abroad in translation—Heinrich Böll (a Nobel laureate) and Günter Grass; some others gaining widespread recognition more recently are Siegfried Lenz, Peter Weiss, Uwe Johnson, Hans Magnus Enzensberger, Walter Kempowski, Peter Handke, and Gabriele Wohmann. During the partition, many East German writers departed to the West, some after a term of imprisonment. Among those of international significance who remained in East Germany were Stefan Heym, Anna Seghers, and Christa Wolf.

The German theatre has long been faced with the dilemma of either relying on the rich repertoire of German classics from the 18th and 19th centuries, along with a restricted number of established 20th-century dramas, as, say, those of Brecht or Carl Zuckmayer, together with contemporary plays in translation from Britain, France, and the United States, or of being bold and innovative. On the one hand, the system of public subsidies and ticket subscriptions favours a steady diet of Goethe, Schiller, Gotthold Ephraim Lessing, Shakespeare, Shaw, Jean Anouilh, Chekhov, and Brecht; but, on the other hand, it also allows for risks to be taken and for the plays of newer dramatists such as Martin Walser and Patrick Süskind to be performed. Small experimental theatres enjoy a lively, if hazardous, existence in the major cities and often closely resemble Germany's long-lived and still lively convention of political cabaret.

Bertolt Brecht in Berlin

In the Democratic Republic, all theatres were owned by the state. The German Theatre (Deutsches Theater) in Berlin reopened in September 1945 and was the first German theatre to perform following the Nazi collapse. The old German National Theatre (Deutsches Nationaltheater) in Weimar was the first to be rebuilt after 1945. Understandably, Berlin dominated theatrical developments, especially because of the work of Brecht at the Theater am Schiffbauerdamm. Given a haven in East Germany—a theatre and a company, along with the political and artistic latitude he required—Brecht was able to produce and perform his own works exclusively. The Berliner Ensemble, as his group was named, possibly commanded more critical and popular attention in the West than on its home ground; his plays are still widely performed in the West, and his dramatic theory, notably the "alienation effect" (*Verfremdungseffekt*), has become a rubric in the canon of the performing arts. After his death in 1956, the theatre continued intact under the direction of, first, his widow, the actress Helene Weigel, and then various successors.

Music and dance. Germany's concert halls are faced with the same dilemma as its theatres, of whether to defer to the public's preference for 18th-, 19th-, and early 20th-century composers or to give contemporary works a better hearing. The works of Hans Werner Henze, whose opera

The Young Lord won international acclaim, joined Bernd Alois Zimmerman and especially Karlheinz Stockhausen as the most significant German composers in the last half of the 20th century. The rich legacy of German music (which until the 20th century is understood to encompass the contributions of Austrian composers) includes the work of German-born English composer George Frideric Handel, Georg Philipp Telemann, and Johann Sebastian Bach, the last of whom dominated the late Baroque period, along with that of Austrians Franz Josef Haydn and Wolfgang Amadeus Mozart from the Classical period and Ludwig van Beethoven and Franz Schubert, who bridged the 18th and 19th centuries and the Classical and Romantic periods, and German romantic composer Felix Mendelssohn. Other significant contributors to the history of German music are Robert and Clara Schumann, Carl Maria von Weber, Johannes Brahms, and Richard Wagner (19th century), Richard Strauss and Austrian Gustav Mahler (late 19th and early 20th century), and Kurt Weill, Paul Hindemith, and Austrian Arnold Schoenberg (20th century).

The Berlin Philharmonic is among the world's leading orchestras, as are the Gewandhaus Orchestra of Leipzig; the Bamberg Symphonic Orchestra; the Bavarian Radio Symphony Orchestra; and the Stuttgart Chamber Orchestra. Moreover, the opera houses of Hamburg (the oldest), Berlin, Cologne, Frankfurt, and Munich are world famous.

Among the Germans who left their mark on rock music are the groups Can, Faust, and Tangerine Dream, whose music was dubbed "Krautrock" by Anglo-American critics, and Kraftwerk, who helped lay the foundation for modern techno music.

Dance

Dance has also been important in Germany. In the 18th century the waltz developed from regional social dances of southern Germany and Austria and gained popularity throughout Europe. Modern dance was embraced as *Ausdruckstanz* ("expressionistic dance"). One of its best-known proponents was Kurt Jooss. Dancer Mary Wigman and her student Hanya Holm had a significant impact on modern dance, particularly in the United States. The Stuttgart Ballet rose to world prominence in the 1960s under its South African-born director John Cranko, and in the 1970s in Wuppertal choreographer Pina Bausch pioneered the influential *Tanztheater* ("dance theatre").

The visual arts. Germany's tradition in the visual arts dates from the reign of Charlemagne. In the 15th and 16th centuries the painters Albrecht Dürer, Lucas Cranach the Elder, Matthias Grünewald, and Hans Holbein the Younger ushered in a golden age of German art, which nevertheless did not develop a definite national character until the mid-18th century. At the turn of the 19th century, Romanticism blossomed, perhaps best exemplified in the work of Caspar David Friedrich. German painters of the 20th century, especially those in groups such as Die Brücke ("The Bridge") and Der Blaue Reiter ("The Blue Rider"), developed a new Expressionist current in European art. Beginning in 1916, Kurt Schwitters, George Grosz, Hannah Höch, and others explored the more theoretical concerns of Dada, while in the 1920s artists such as Otto Dix and photographer August Sander worked in the realistic, socially critical style known as Neue Sachlichkeit ("New Objectivity").

These and other developments came to a halt in 1933 with the rise of the National Socialists, who considered such art "degenerate." After World War II, German art struggled to regain a sense of direction, challenged by the emigration of many important German artists. In East Germany a form of Socialist Realism dominated; in West Germany artists experimented with avant-garde movements such as Abstract Expressionism, Pop art, minimalism, and Op art. Beginning in the 1960s, Joseph Beuys created sculpture, performance art, and installation art that challenged the very definition of "high art," while Gerhard Richter gained fame in the 1970s for his paintings based on photographs. German painters such as Georg Baselitz, Anselm Kiefer, and Sigmar Polke were at the centre of the art world when Neo-Expressionism became dominant internationally in the 1980s. At the turn of the 21st century, German photographers Wolfgang Tillmans, Bernd and Hilla Becher, Thomas Struth, and Andreas Gursky won international art prizes and prominence.

Contemporary German architecture—indeed world architecture—is very much the creature of the Bauhaus school, founded in Weimar in the 1920s by Walter Gropius and directed for a time by Ludwig Mies van der Rohe. After World War II the dogmas of the Bauhaus school—the insistence on strict harmony of style with function, and on the intrinsic beauty of materials, as well as a puritan disdain of decorativeness—were widely applied. Yet in the 1960s and '70s, the stark Bauhaus style began to yield to the more free-ranging postmodernism, which took as its precept "not just function but fiction as well."

The Bauhaus style

Reflecting the Soviet influence, buildings in eastern Germany differ from those in western Germany in the immensity of their proportions. The major showpieces in eastern Berlin—the government buildings, apartment blocks, hotels, and public spaces along Unter den Linden, Marx-Engels-Platz, Alexanderplatz, Karl-Marx-Allee, and Leipziger-Strasse—and their exaggerated decorations testify to a propensity for sheer vastness. After unification the long-deserted Potsdamer Platz in Berlin came alive with the construction of an array of public and private buildings by internationally renowned architects.

Film. Before World War II Germany helped set the pace in motion pictures. Directors such as Fritz Lang, Ernst Lubitsch, F.W. Murnau, and G.W. Pabst had virtually defined cinematic art in the 1920s and early '30s. In the Nazi period, however, most noteworthy German filmmakers, including Lang, Lubitsch, Murnau, and Billy Wilder, relocated to Hollywood, as did actors such as Marlene Dietrich and Conrad Veidt. Documentarian Leni Riefenstahl (*Triumph of the Will,* 1935), however, remained in Germany.

After World War II the studios of the giant UFA (Universum-Film-Aktiengesellschaft), which were in East German hands, continued as a government-owned enterprise known as DEFA (Deutsche Film-Akademie), noted for animation, documentaries, and feature films. Germany's reunification in 1990, however, cut short attempts to forge a national cinematic identity in eastern Germany. In West Germany a group of young filmmakers, first organized at the Oberhausen Film Festival in 1962, established *das neue Kino,* or the New German Cinema; though they had little commercial success outside of Germany, the films of directors such as Rainer Werner Fassbinder, Volcker Schlondorff, Margarethe von Trotta, Werner Herzog, and Wim Wenders won international critical acclaim, as did the work of actors such as Klaus Kinski, Bruno Ganz, and Hanna Schygulla. Some German filmmakers also found success in the United States, including Roland Emerich, who established himself in the action-adventure genre, and Wolfgang Petersen (*Das Boot,* 1982), who became one of Hollywood's more noted practitioners of the thriller. Though at the beginning of the 21st century there was no shortage of German filmmaking talent, few German directors were producing German films.

Arts festivals. Most major cities and scores of small towns sponsor festivals. Among the best known is the Bayreuth Festival celebrating the works of Richard Wagner (who founded the festival himself in 1876). The oldest German festival is the Passion play, first held in 1634 and now held every 10 years in Oberammergau (Bavaria).

The Bayreuth Festival

Berlin has five major festivals: the Berliner Festwochen (Berlin Festival Weeks) in September and October, with five weeks of musical events; the Berliner Jazzfest in November; the Berlin International Film Festival in February; the Theatertreffen Berlin ("Berlin Theatre Meeting"), featuring productions from throughout the German-speaking world; and the Karneval der Kulturen ("Carnival of Cultures"), a festival of world cultures. Munich has an opera festival in July and August, with emphasis on Richard Strauss. Festivals in Würzburg and Augsburg celebrate Mozart. Ansbach has a Bach festival, and Bonn has one honouring Beethoven. Other noteworthy events include Documenta, an arts festival held every five years in Kassel, the International May Festival in Wiesbaden, and the Festival of Contemporary Music in Donaueschingen. Expo 2000, Germany's first world's fair, was held in Hanover.

Publishing. Germany has some 2,000 publishing houses, and more than 50,000 titles reach the public each year, a production surpassed only by the United States. Germany traditionally was home to small and medium-size publishing houses. However, the Gütersloh-based Bertelsmann group, a multinational conglomerate, is one of the world's largest publishers. Book publishing is not centred in a single city but is concentrated fairly evenly in Berlin, Hamburg, and the regional metropolises of Cologne, Frankfurt, Stuttgart, and Munich. Leipzig, prewar Germany's major publishing city, shared with East Berlin the major publishing houses of East Germany. Gotha in Thuringia is renowned for the production of maps and atlases.

The press. Germans are voracious readers of newspapers and periodicals. Freedom of the press is guaranteed under the Basic Law, and the economic state of Germany's several hundred newspapers and thousands of periodicals is enviably healthy. Most major cities support two or more daily newspapers, and few towns of any size are without their own daily newspaper. During the 1990s most German newspapers and periodicals published daily or weekly editions on the Internet, enabling access far beyond their traditional print circulation.

The press is free of government control, no newspaper is owned by a political party, and only about 10 percent of newspapers overtly support a political party, though most offer a distinctly political point of view. Laws restrict the total circulation of newspapers or magazines that can be controlled by one publisher or group. The Bundeskartellamt (Federal Cartel Office) oversees German industry (including the media) to ensure against a company abusing its dominant position within a particular industry. Although newspaper and periodical ownership cannot be the monopoly of any one ownership, Axel Springer Verlag AG controls a significant share of the market. The German Press Council, established in 1956, sets out guidelines and investigates complaints against the press.

Newspapers

A national press exists on one level in the form of *Süddeutsche Zeitung* (Munich), *Die Welt* (Berlin), and the *Frankfurter Allgemeine Zeitung,* together with regional newspapers (*e.g.,* the *Stuttgarter Zeitung* and the *Frankfurter Rundschau*), which also command international circulation and respect. The nationally circulated tabloid *Bild* (Hamburg), which has the largest readership of any paper, publishes several regional editions.

Berlin has many daily newspapers, including the liberal *Der Tagesspiegel* and the conservative *Berliner Morgenpost.* Originally published in East Germany but acquired by western press interests after unification, the *Berliner Zeitung* has become the city's preeminent newspaper. Other leading newspapers of the former East Germany were also bought by western publishers.

The major editions of German newspapers are published on Saturdays. A lively Sunday press complements the daily newspapers, providing an overview, perspective, and interpretation of major news developments as well as political comment and artistic criticism; the most prestigious and influential of these is *Die Zeit* (Hamburg). Others include the venerable *Rheinischer Merkur* (Bonn), the *Bayernkurier* (Munich), and the *Deutsches Allgemeines Sonntagsblatt* (Hamburg). Sunday counterparts of the major dailies, *Welt am Sonntag* and *Bild am Sonntag,* are run virtually as separate newspapers.

The genre of the *Illustrierte* (pictorial) dominates the German magazine market. Some of these popular weekly glossies, such as *Stern* and *Bunte,* carry features, including investigative reporting, of a high calibre; others cater to the public thirst for the escapades of celebrities, bizarre crime, tales of gracious living, and sundry escapist topics. There is also a wealth of specialized journals and quality business-oriented magazines. A special niche is occupied by the weekly newsmagazine *Der Spiegel,* a journalistic power in its own right. Founded in the aftermath of World War II, it has shaped public opinion in Germany through its editorial posture as the skeptical, nonaligned observer and guardian of the public conscience.

Der Spiegel

Broadcasting. Although German radio and television are not state-controlled, only public corporations were permitted to broadcast until the mid-1980s, when a dual system was established. Still, in 1986 the Federal Constitutional Court held that the public corporations comprised the "basic supply" of news and entertainment, and commercial outlets were only a "supplementary supply." The Federal Ministry of Post and Telecommunications oversees public broadcasting, though the public corporations have great freedom in establishing their own broadcasting policies, and German television (more than radio) enjoys remarkable latitude and independence. Support is provided by fees paid by the owners of radios and television sets.

Public radio and television are arranged along national and regional lines, with a number of regional corporations that offered two to four radio programming schedules combining to form one evening television offering, ARD (Arbeitsgemeinschaft der Öffentlich-Rechtlichen Rundfunkanstalten Deutschlands). This is complemented by a second television network, ZDF (Zweites Deutsches Fernsehen), which is based in Mainz. A third channel is operated by ARD but is organized and broadcast regionally, with special emphasis placed on local and regional events, along with educational, informational, and fine arts programs. The uneven quality of entertainment in both radio and television is offset by the high-quality news coverage and political and social reporting that makes the German public one of the best-informed of any country.

Two radio stations—Deutschland Radio and Deutsche Welle—are publicly operated to provide a comprehensive German perspective of events; Deutsche Welle is beamed to Europe and overseas. There are also several regional public radio stations and some 200 private radio stations that are regionally and locally focused.

Cable television has been greatly expanded. Supplied by satellite transmissions, the cable networks offer extensive programming from public and commercial television in Germany and from abroad.

During the years of partition, viewers in East Germany could freely receive radio and television broadcasts from West Germany and from West Berlin, with the result that the public in East Germany kept current on news from the West. The broadcasting facilities in the former East Germany were reorganized along lines of the western states—*i.e.,* each of the new states has its own regional stations.

Sports and recreation. In the early 19th century, coincident with the rise of nationalism, Friedrich Ludwig Jahn founded the turnverein, a gymnastics club, and invented several of the disciplines that are now part of the Olympic gymnastics program. At the same time, Johann Christoph Friedrich GutsMuths helped bring physical education to the forefront of German education. Ideals of health—*Gesund*—and athletic prowess became important components of the German conception of self and were later critical to the Nazi conception of the ideal German.

The turnverein

In 1936 Berlin hosted the Summer Olympics, which Adolf Hitler transformed into a stage to promote Nazi ideals, though this effort was thwarted somewhat by key victories by African American sprinter and long jumper Jesse Owens and other "non-Aryan" athletes. The 1972 Olympics, held in Munich, were also marred when Palestinian terrorists took hostage and killed 11 members of the Israeli team. Following partition, East and West German athletes competed on the same national team from 1956 to 1964 but split their teams for the next six Olympiads, joining the front lines of Cold War athletic competition. Germany's Olympic teams—East, West, and unified—have dominated many events, especially swimming, in which East German Kornelia Ender won 26 gold medals.

Football (soccer) is a passion for many Germans, and the national Bundesliga is among the world's most respected professional football leagues. Moreover, West Germany won the World Cup in 1974 and 1990. Germans have also excelled in professional tennis, golf, and basketball.

Leisure is a major pursuit in German life. As did workers in most Western industrial countries, Germans once toiled under an unrestricted six-day workweek, often 10 hours a day with breaks only on Sundays and for major feasts and festivals. The German workweek is now 40 hours or less and German workers receive three to six weeks or more of paid vacation time. As a result, Germans have more leisure time than workers of most Western countries. This

abundance of leisure has become a national preoccupation.

To a surprising extent in a country so highly industrialized, ancient feasts and practices are still widely observed in both Roman Catholic and Protestant areas, especially pre-Lenten celebrations known as *Fasching* in the southern regions or *Karneval* (carnival) in the Rhineland. In addition to the major religious festivals—Easter, Christmas, Whitsun (which are also national holidays), and, in Roman Catholic districts, Corpus Christi Day and Assumption—there are many local celebrations—wine, beer, harvest, hunting, and historical festivals—the so-called *Volksfeste* that are deeply rooted in German custom, the most notable being Munich's Oktoberfest, held each September.

Oktoberfest

Although the traditional observances continue unabated, Germans have adopted many more modern forms of recreation, amusement, and relaxation. Travel has become the favoured pastime of a majority of Germans. More than half of all adults take at least one annual trip for pleasure, and a great number take more than one. Many take both a winter and summer vacation. Older persons often take a paid *Kur* at a spa to rest and recuperate, in addition to a holiday trip for pleasure. More Germans take leisure trips abroad than the citizens of any other country.

The weekend, a latecomer in Germany, has become firmly established. Private recreation typically includes spectator amusements and sports—especially football, active sports and physical exercise, automobile excursions, the pursuit of hobbies, visits with friends and family, and the long-favoured German pastime of walking or hiking.

It is estimated that about one-fifth of a household's income in Germany's western regions is spent on leisure expenses. Many institutions, including the government, local communities, schools, churches, and companies, encourage citizens to channel their free time into useful, rewarding, and healthful pursuits by providing physical facilities, impetus, and other prerequisites for public recreation. Leisure in Germany is now regarded—much like education and vocational training, housing, a job, good public transportation, health and disability insurance, and pensions—as an entitlement and a valuable adjunct of social policy.

In East Germany, leisure activity was arranged very differently. The state ideal was group leisure, holidays, and travel, often organized by the workplace or a youth organization. Cheap holiday facilities were available, especially along the resorts of the Baltic coast. Residents of East Germany were free to travel privately to any of the Warsaw Pact countries and sometimes to Yugoslavia. It is worth noting that the fall of the communist bloc was precipitated in part by the large number of East Germans who, while visiting Hungary, crossed unimpeded into Austria when the Hungarian government opened its borders.

CULTURAL INSTITUTIONS

Many organizations, maintained entirely or in part by public funds, are devoted to acquainting the world with the culture, life, and language of the German peoples and familiarizing Germans with foreign cultures. Cultural representation abroad is abundantly maintained in the advanced industrial countries of the West and in eastern Europe, but special emphasis is placed on fostering educational and cultural ties with less-developed countries. For them Germany is not only a major lender of technological skill and capital for developing resources but also an important centre for the education of students from these countries.

The Goethe-Institut Inter Nationes (formerly the Goethe-Institut of Munich), founded in 1951, has some 140 branches in over 70 countries. It operates schools in Germany and abroad that offer instruction in the German language. It also maintains lending libraries and audiovisual centres and sponsors exhibits, film programs, musical and theatrical events, and lectures by prominent personalities.

Museums and galleries. Germany has some 2,000 museums of all descriptions, housing some of the world's great collections of painting and sculpture as well as of archaeological and scientific displays. Notable museums and galleries include the museums of the Prussian Cultural Property Foundation in Berlin—*i.e.,* the Pergamon Museum with its vast collection of Classical and Middle Eastern antiquities, located on the "Museum Island" in the River Spree, together with the Old (Altes) Museum, the New (Neues) Museum, the National Gallery (National-galerie), and the Bode Museum—the Zwinger Museum and Picture Gallery in Dresden, the Bavarian State Picture Galleries and the Deutsches Museum in Munich, the Germanic National Museum in Nürnberg, the Roman-Germanic Central Museum in Mainz, the Senckenberg Museum of natural science in Frankfurt am Main, and the State Gallery in Stuttgart. Some museums are highly specialized, devoted to a single artist, school, or genre, but many combine natural science and fine arts. Among the many ethnological museums are the Linden Museum in Stuttgart, the East German Gallery Museum in Regensburg, and the Ethnological Museum in Berlin-Dahlem. Important art treasures are scattered in the scores of smaller museums, libraries and archives, castles, cathedrals, churches, and monasteries throughout the country. The Berlin-Dahlem Botanical Garden and Botanical Museum, founded in the 17th century, is German's oldest botanical garden. Specialized exhibits, from the 500th anniversary celebration of Albrecht Dürer in Nürnberg (1971) to the Cézanne (1993) and Renoir (1996) exhibitions at Tübingen Kunsthalle, have attracted visitors from throughout the world.

Museum Island

Libraries. Among Germany's great libraries are the Bavarian State Library in Munich (the largest) and the library of the Prussian Cultural Property Foundation (formerly the Prussian State Library) and National Library in Berlin. The German Library at Frankfurt am Main is the country's library of deposit and bibliographic centre. The Technical Library at Hannover is Germany's most important library for science and technology and for translated works in the fields of science and engineering. The great university libraries at Heidelberg, Cologne, Göttingen, Leipzig, Tübingen, and Munich are complemented by scores of other good university libraries. A wealth of manuscripts, early printed works, and documents from the Middle Ages to the present are dispersed in smaller collections. There is also an extensive system of lending libraries operated by the states, the municipalities, the library associations of the Roman Catholic and Evangelical churches, and other public associations and institutes. (G.H.K./W.H.Be.)

For statistical data on the land and people of Germany, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Ancient history

The German peoples are defined by the common language group to which they belong. German history thus originates with the so-called first sound shift (or Grimm's law), which turned a Proto-Indo-European dialect into a new Germanic language group. The Proto-Indo-European consonants *p, t,* and *k* became the Proto-Germanic *f, þ* (*th*), and *x* (*h*), and the Proto-Indo-European *b, d,* and *g* became Proto-Germanic *p, t,* and *k.* The historical context of the shift is difficult to identify because it is impossible to date it conclusively. Clearly the people who came to speak Germanic must have been isolated from other Indo-Europeans for some time, but it is not obvious which archaeological culture might represent the period of the shift. One possibility is the so-called Northern Bronze Age, centred in northern Germany and Scandinavia, that flourished between about 1700 and 450 BC. Alternatives would be one of the early Iron Age cultures of the same region (*e.g.,* Wessenstadt [800–600 BC], or Jastorf [600–300 BC]).

The Germanic consonant shift

Solid historical information begins in about 50 BC when Julius Caesar's Gallic Wars brought him into contact with Germans as well as Celts. He did cross the Rhine in 55 and 53 BC, but the province of Gaul he created used

the river as a boundary and most Germans lived beyond it. Direct Roman attacks on German tribes began again under Nero Claudius Drusus Germanicus, who pushed across the Rhine in 12–9 BC, while other Roman forces assaulted Germanic tribes through the middle Danube (in modern Austria and Hungary). Fierce fighting in both areas, and the famous victory of the German Arminius in the Teutoburger Forest in AD 9 (when three Roman legions were massacred), showed that conquering these tribes would require too much effort. The Roman frontier thus stabilized on the Rhine and Danube rivers, although sporadic campaigns (notably under Domitian in AD 83 and 88) extended control over Frisia in the north and some lands east of the confluence of the Rhine and the Danube.

Both archaeology and Caesar's own account of his wars show that German tribes then lived on both sides of the Rhine. The Romans also met Germans on the middle Danube. In fact, broadly similar archaeological cultures from this period stretch across central Europe from the Rhine to the Vistula River (in modern Poland), so that Germanic peoples probably dominated all of these areas. Germanic cultures extended from Scandinavia as far south as the Carpathians. These Germans led a largely settled agricultural existence. They practiced mixed farming, lived in wooden houses (working mainly in wood), did not have the potter's wheel, were nonliterate, and did not use money. The marshy lowlands of northern Europe preserve otherwise perishable wooden objects, leather goods, and clothing and shed much light on the Germanic way of life. These bogs were also used for ritual sacrifice and execution, and some 700 "bog people" have been recovered. Their remains are so well preserved that even dietary patterns can be established; the staple was a gruel made of many kinds of seeds and weeds.

Social structure

Clear evidence of social differentiation appears in these cultures. Richly furnished burials (containing rich jewelry and sometimes weapons) have been uncovered in many areas, showing that a wealthy warrior-prince class was developing. These chiefs became a standard feature of Germanic society, and archaeologists have uncovered the halls where they feasted their retainers, an activity described in the Anglo-Saxon poem *Beowulf*. This warrior elite followed the cult of a war god (Tiu or Wodan). Tacitus describes in the *Germania* how in AD 59 the Hermunduri, in fulfillment of their vows, sacrificed defeated Chatti to this god. This elite was also the basis of political organization. The Germans were divided into numerous tribes, which were also united in leagues centred on the worship of particular cults. These cults were probably created by one locally dominant tribe and changed over time. Tribes belonging to such leagues came together for an annual festival, when weapons were laid aside. Apart from worship, these were also times for economic activity, social interaction, and settling disputes.

COEXISTENCE WITH ROME TO AD 350

After Rome had established its frontiers, commercial and cultural contacts were as important as direct conflict. Although it was heavily fortified, the frontier was never a barrier to trade or people. In about AD 50, tribes beside the Rhine learned to use Roman money. Germanic graves—at least the richer ones—began to include Roman luxury imports such as fine pottery, glass, and metalwork. In return, raw materials such as amber and leather, and many slaves, went back across the frontier. Germans also served in Roman armies.

Border raiding was endemic, and periodically there were much larger disturbances. In about AD 150 the Marcomanni, a Germanic tribe, moved south into the middle Danube region, and they even invaded Italy in 167. The emperor Marcus Aurelius and his son spent the next 20 years curbing their inroads, and archaeology shows that the wars were highly destructive. This migration was one manifestation of a broader problem, for between about 150 and 200 a whole series of Germanic groups moved south along the river valleys of central and eastern Europe. These migrations resulted in great violence along the entire frontier during the 3rd century. Parts of Gaul suffered greatly, and Goths ravaged the Danube region, even killing the emperor Decius in 251. Yet intensive campaigns brought the Germanic tribes back under control, so that by about 280 stability had returned to the Rhine and Danube. The Roman army and an alliance system involving, among others, Franks, Alemanni, and Goths maintained the frontier until about 370.

Transformation of the Germanic world

In the meantime the Germanic world was being transformed. For the balance of power in Europe, most important was the rise of larger and more cohesive Germanic political units, at least among the Germans living on the borders of the empire. This was largely a response to the military threat from Rome. Despite their occasional successes, more Germans than Romans had been killed in 3rd-century conflicts, and the Germans had learned that larger groups were more likely to survive. In the 4th century there were two powerful Germanic confederations: the Alemanni on the Rhine and the Goths on the Danube. These were sustained by the military elite whose control over their fellow tribesmen continued to increase. Other contacts with the empire had fewer political effects. In the 3rd century Germanic cultures began to master the potter's wheel, and there is evidence of improved farming techniques; both were adopted from Rome. The empire was also partly responsible for the Germans' first steps toward literary culture. Gothic, the oldest literary Germanic language, was created in about AD 350 by Ulfilas, a Roman-sponsored missionary, in order to translate the Bible.

THE MIGRATION PERIOD

The situation was revolutionized by nomadic (non-German) Hunnic horsemen from the east who pushed Germanic peoples into the Roman Empire in several waves. In 376, first of all, Visigoths were admitted by the emperor Valens, but only unwillingly, and confrontation followed. Two years later they killed Valens, winning a famous victory at Hadrianople, though by 382 they had been subdued. Yet, as the Huns moved west, Rome's frontiers came under increasing pressure, and further large incursions (Germanic and others) occurred in 386, 395, 405, and 406. Some of the invaders were defeated, but Germanic Vandals and Suebi established themselves in North Africa and Spain, and the Visigoths exploited the disorders to rebel. Marching to Italy, they demanded better terms, and, when these were not forthcoming, they sacked Rome on Aug. 24, 410.

The Roman Empire nevertheless remained an important power in Europe, having both troops and funds. Hence Germans on the run from the Huns were anxious to make peace; even the Visigoths accepted a settlement in Gaul in 418. Since Germanic groups had no sense of Germanic nationalism, they could be played off against one another; thus Vandals were savaged by the Visigoths between 416 and 418. Until about 450 fear of the Huns meant that the empire could, in moments of crisis, mobilize at least Visigoths, Burgundians (received into Gaul after being defeated by the Huns in 439), and Franks for its defense. Soon after Attila's death in 453, however, the Hun empire collapsed, and Rome lost this diplomatic weapon. It also suffered a progressive loss of revenue as territories were either occupied, or—like Britain—declared themselves independent. The balance thus swung further in the Germans' favour, and they eventually declared themselves independent. In the 470s a Visigothic kingdom emerged in southwestern Gaul and a Burgundian kingdom in southeastern Gaul, and Clovis created a Frankish kingdom in the north. The Vandals already controlled North Africa and the Suebi part of Spain. Ostrogoths freed by the collapse of the Hun empire conquered Italy between 489 and 493, and Gepidae and Lombard kingdoms dominated the Danube.

The rise of independent kingdoms

These Germanic successor states brought the Roman Empire in western Europe to an end. The empire could have resisted any of them singly, but the Hun invasions had pushed too many Germans across the frontier too quickly. Battles, however, were the exception. More often the empire unwillingly, but peacefully, granted areas for settlement, and as Rome became increasingly less powerful, the local Roman provincial population looked to the newly settled Germans for protection. This period thus continued the transformation of the Germanic world. Be-

cause of the danger from Rome and the Huns, Germanic political units again increased in size. The new Germanic groups also fell under the influence of the Roman populations they came to rule. Literate, educated Romans enabled German kings systematically to raise taxation and expand their legal powers. The successor states to the Roman Empire were thus a fusion of Germanic military power and the administrative know-how of Roman provincial aristocrats. Transformation was complete when Germanic warrior and Roman provincial elites quickly intermarried, bringing into being a new aristocracy that was to shape medieval Europe. (P.J.He.)

Merovingians and Carolingians

When the western Roman Empire ended in 476, the Germanic tribes west of the Rhine were not politically united. The west Germanic tribes, however, spoke dialects of a common language and shared social and political traditions. The traditions of these tribes had been influenced by centuries of contact with the Roman world, both as federated troops within the empire and as participants in the broader political and economic network that extended beyond the Roman frontier. In particular a strongly military structure of social organization, under the direction of commanders termed kings or dukes, had developed among the federated tribes within the empire and spread to tribes living outside the empire proper. Likewise, the Ostrogothic kings in Italy extended their influence over much of the Germanic world north of the Alps.

MEROVINGIAN GERMANY

The beginnings of Frankish domination

The Franks, settled in Romanized Gaul and western Germany, rejected Ostrogothic leadership and began to expand their kingdom eastward. Clovis' conversion to orthodox Christianity was an overt challenge to Ostrogothic hegemony to the east and to Visigothic control to the south. He and his successors, particularly Theodebert I (reigned 534–548), brought much of what would later constitute Germany under Frankish control, including the Thuringians of central Germany and the Alemanni and Bavarians of the south. Generally these heterogenous groups were given a law code including Frankish and local traditions and were governed by a duke of mixed Frankish and indigenous background who represented the Frankish king. In times of strong central leadership, as under Dagobert I (629–639), this leadership could have real effect. At other times, when the Frankish realm was badly divided or embroiled in civil wars, local dukes enjoyed great autonomy. This was particularly true of the Bavarian Agilolfings, who were closely related to the Lombard royal family of Italy and who by the 8th century enjoyed virtual royal status. In the north the Frisians and Saxons remained independent of Frankish control into the 8th century, preserving their own political and social structures and remaining for the most part pagan. In areas under Frankish lordship, Christianity made considerable progress through the efforts of native Raetians in the Alpine regions, of wandering Irish missionaries, and of transplanted Frankish aristocrats who supported monastic foundations.

THE RISE OF THE CAROLINGIANS AND BONIFACE

By the end of the 7th century, Merovingian rule throughout the Frankish world had been destroyed by powerful, competing aristocratic clans, each claiming autonomy and hoping to establish hegemony over the Frankish realm. The clan that finally succeeded was the Carolingian, which drew its strength from extensive estates and loyal aristocratic supporters in the lands between the Meuse and the Rhine. The Carolingians ruled the kingdom from the 730s, although they did not acquire the royal title until 751. They consolidated effective control over the Frankish heartland and the duchies east of the Rhine. Consolidating control over these duchies was facilitated by supporting the missionary activities of churchmen closely allied with Roman hierarchical forms of ecclesiastical organization that favoured political centralization, at the expense of indigenous and Irish ecclesiastical structures. The pattern of Frankish penetration was always the same. Small communities or churches were settled on land newly won from forest or marsh, granted them by their Carolingian protectors. Thus, from Frisia in the north to Bavaria in the south, religious, economic, and political penetration went hand in hand. A distinguished part was played by Anglo-Saxon missionaries, who linked the Frankish world not only with the high culture of the churches (notably York) from which they set out but also with Rome, the ultimate source of their inspiration. Chief among them were St. Willibrord (*c.* 658–739), who worked as a missionary and Frankish agent among the Frisians and later the Thuringians, and St. Boniface (*c.* 675–754), who created a German church structure. Supported by Charles Martel against both the aristocratic Frankish clergy and the preexisting clergy of the Germanic regions, many of whom feared growing Carolingian and Roman control, Boniface led missions into Franconia, Thuringia, and Bavaria, where he founded or reorganized diocesan organization on a Roman model. In 742 he played a large part in the first council of the new German church. By the time of his death at the hands of northern Frisians, all of the continental Germanic peoples except the Saxons were well on the way toward integration into a Roman-Frankish ecclesiastical structure.

Creation of the German church

CHARLEMAGNE

Charlemagne built on the foundations laid by Charles Martel and Boniface. Contemporary writers were vastly impressed by Charlemagne's political campaigns to destroy the autonomy of Bavaria and his equally determined Saxon military campaigns. Under their Agilolfing dukes, who had at times led the opposition to the rising Carolingians, the Bavarians had developed an independent, southward-looking state that had close contacts with Lombard Italy and peaceful relations with the Avar kingdom to the east. Charlemagne's conquest of the Lombards in 774 left Bavaria isolated, and in 788 Charlemagne succeeded in deposing the last Agilolfing duke, Tassilo III, and replacing him with a trusted agent. Thereafter, Charlemagne used Bavaria as the staging ground for a series of successful wars that ultimately destroyed the Avar kingdom.

The subjugation of the north proved much more difficult than that of the south. In the wake of the missionaries, Frankish counts and other officials moved into northeastern Frisia, raising contingents for the royal host and doing the other business of secular government. As for the Rhineland, the richer it grew the more necessary it became to protect its hinterland, Franconia (Hesse) and Thuringia, from Saxon raids. Because there was no natural barrier behind which to hold the Saxons, this was a difficult task.

Unlike the Bavarians, the Saxons were not politically united. Their independent *edhelingi* (nobles) lived on estates among forest clearings, dominating the *frilingi* (freemen), *lazzi* (half-free), and unfree members of Saxon society and leading raids into the rich Frankish world. Thus each of Charlemagne's punitive expeditions bit deeper into the heart of Saxony, leaving behind bitter memories of forced conversions, deportations, and massacres. These raids were inspired by religious as well as political zeal; Charlemagne tried to break Saxon resistance both to Christianity and to Frankish dominance with fire and sword. Still, the decentralized nature of Saxon society made ultimate conquest extremely difficult. Whenever the Frankish army was occupied elsewhere, the Saxons could be counted on to revolt, to slaughter Frankish officials and priests, and to raid as far westward as they could. Charlemagne in turn would punish the offending tribes and garrison the defense points abandoned by the Saxons. In time resistance to the Franks gave the Saxons a kind of unity under the leadership of Widukind, who succeeded longer than any other leader in holding together a majority of chieftains in armed resistance to the Franks. Ultimately internal feuding led to the capitulation even of Widukind. He surrendered, was baptized, and, like Tassilo, was imprisoned in a monastery for the remainder of his life. Saxony had been savagely repressed, as is reflected in the *Capitulatio de partibus Saxoniae* (*c.* 785; "Capitulary for the Saxon Regions") and the *Capitulare Saxonicum* (797; "Capitulary of the Saxons"), both measures intended to force the submission of the Saxons to the Franks and

The subjugation of Saxony

to Christianity. Loyal Frankish churchmen and aristocrats were introduced to secure and pacify the region. Although northern regions that enjoyed Danish support remained outside of Frankish control, most of Saxony gradually moved into the united Frankish realm.

THE EMERGENCE OF GERMANY

The kingdom of Louis the German. Charlemagne made no attempt to rule his vast empire in a unified manner and was content to leave each Germanic region largely in the hands of his counts and bishops. His son Louis I (Louis the Pious) was not unpopular with his Germanic subjects; on two occasions he owed his restoration to power largely to their support, but his primary focus was on the Romance-speaking regions of the empire. In 825 he entrusted his son Louis the German (804–876) with the government of Bavaria, whence he was gradually to extend his power over all of Carolingian Germany. This was the first time that the German peoples had had a ruler whose authority was confined to their own lands. Although his ambitions extended throughout the whole Carolingian world, Louis cultivated German language and literature for the first time as a self-conscious cultural and political identity. Under his patronage the Gospels were translated into Germanic dialects and the first attempts at writing Germanic poetry with Christian and traditional themes were undertaken. Louis himself was described as the Frankish king of the eastern kingdom. Much of Louis' reign was taken up with campaigns against neighbouring Slavs, and this external focus maintained stability and royal authority over the aristocracy. By 870 Louis' dominions reached almost the boundaries of medieval Germany. On the east they were bordered by the Elbe and the Bohemian mountains; on the west, beyond the Rhine, they included the districts afterward known as Alsace and Lorraine. Ecclesiastically, they included the provinces of Mainz, Trier, Cologne, Salzburg, and Bremen. Although the close kinship and rivalries of the descendants of Charlemagne still united east and west Francia, the eastern region was taking on the identity of Germany and the west was emerging as France.

Germany and France

Louis' long reign had accustomed the Germanic peoples of his kingdom to a certain unity. After his death the kingdom was divided among his three sons in keeping with Carolingian tradition, but the deaths of two of them, in 880 and 882, restored its unity under Charles the Fat. The ceaseless external blows from Danes, Saracens, and Magyars that fell upon the Carolingian world in the later 9th and 10th centuries weakened the kingdom's cohesion, however. Not only the Carolingians themselves but also their followers were prepared to take advantage of one another, to compromise with the enemy, and to carve out even more dominions from one another's lands. Incompetence in mounting effective resistance to invaders led to revolts and civil wars, as for instance in 887 when Arnulf, an illegitimate son of Louis the German's son Carloman, led an army of Bavarians in a successful revolt against his uncle, Charles the Fat. Arnulf, however, was not equally successful in defending his eastern possessions. After his death in 899, the German kingdom came under the nominal rule of his young son, Louis the Child, and in the absence of strong military leadership it became the prey of the Magyar horsemen and other invaders from the east.

Rise of the duchies. Because the Carolingians themselves were unable to provide effective defense for the whole kingdom, military command and the political and economic power necessary to support it necessarily devolved on local leaders whose regions were attacked. The inevitable result was the decentralization and decay of royal authority to the profit of the regional dukes. Contrary to popular opinion, these dukes were not appointed by the peoples concerned, nor were they descendants of the tribal chieftains of the postmigration period. The so-called *Stammesherzogtümer* (stem or tribal duchies) were new political and, ultimately, social units. Their dukes were Carolingian counts, part of the international "imperial aristocracy" of the Carolingians, who took the initiative in organizing defense on a local basis, without thereby seeking to shake men's loyalty to the Carolingians. All the same, their initial success established them in the hearts of those whom they protected. In Saxony the Liudolfings, descendants of military commanders first established by Louis the German, achieved spectacular successes against the Slavs, Normans, and Magyars. In Franconia the Konradings rose to prominence over this largely Frankish region with the assistance of Arnulf but became largely independent during the minority of his son. Thuringia fell increasingly under the protection and lordship of the Liudolfings. In Swabia (Alemannia) the Hunfridinger and Erchanger clans, originally established by Carolingian kings, disputed control with each other and with regional ecclesiastical lords. Similarly the Luitpoldingers, originally named as Carolingian commanders, became dukes of Bavaria. Throughout the kingdom the only force for preserving unity remained the church, but the threat of external foes led to the secularization of much monastic land with episcopal approval, further strengthening the power of the dukes.

The structural transformation of the "imperial aristocracy" to a local elite was accompanied by an increasingly particularist, dynastically oriented aristocratic society that was bound together through ties of vassalage and exercised personal lordship over the free and half-free populations of the regions. This process did not advance as far in Germany as it did in France. Everywhere German society remained closer to older, regional varieties of social organization as well as to traditions of Carolingian ecclesiastical and comital government. (J.M.W.-H./P.J.G.)

Changes in social organization

Germany from 911 to 1250

THE 10TH AND 11TH CENTURIES

Conrad I (911–918). When in 911 Louis the Child, last of the East Frankish Carolingians, died without leaving a male heir, it seemed quite possible that his kingdom would break into pieces. In at least three of the four stem lands, Bavaria, Saxony, and Franconia, the ducal families were established in the leadership of their tribes; in Swabia (Alemannia) two houses were still fighting for hegemony. Only the church, fearing for its endowments, had an obvious interest in the future of the monarchy, its ancient protector. Against the growing authority of the dukes and the deep differences in dialect, in customs, and in social structure between the tribes there stood only the Carolingian tradition of kingship; but, with Charles the Simple as holder of the West Frankish kingdom, its future was uncertain and not very hopeful. Only the Lotharingians put their faith in the ancient line and did homage to Charles, its sole reigning representative. The other component parts of the East Frankish kingdom did not follow suit.

One can only guess at the motives of the Saxon and Frankish tribal hosts who on Nov. 10, 911, elected Conrad, duke of the Franks, as their king at Forchheim in Franconia. At the opening of the 10th century the Germanic peoples in the lands east of the Rhine and west of the Elbe, the Saale, and the Bohemian forest—as rudimentary and as thinly spread as their settlements were—had to face even more primitive and pagan races pressing in from farther east, especially the Magyars. The Saxons, headed by their duke Otto of the house of the Liudolfings, were threatened by more enemies on their frontiers than any other tribe; Danes, Slavs, and Magyars simultaneously harassed their homeland. A king who commanded resources farther west, in Franconia, might therefore prove to be of help to Saxony. The Rhenish Franks, on the other hand, did not wish to abdicate from their position as the leading and kingmaking people, which gave them many material advantages.

Conrad of Franconia, elected by Franks and Saxons, was soon recognized also by Arnulf, duke of Bavaria, and by the Swabian clans. In descent, honours, and wealth, however, Conrad was no more than the equal of the dukes who had accepted him as king. To gain a lead over them, to found a new royal house, and to acquire those wonder-working attributes that the Germans venerated in their rulers long after they had been converted to Christianity, he had yet to prove himself able, lucky, and successful. In this period, political affairs became the monopoly of the German kings and a few score families of great magnates.



Germany in the 10th and 11th centuries.

The reason for this concentration of power was that, at the very foundation of the German kingdom, circumstances had long favoured those men whom birth, wealth, and military success had raised well above the ranks of the ordinary free members of their tribe. Their estates were cultivated in the main by half-free peasants—slaves who had risen or freemen who had sunk. The holdings of these dependents fell under the power of the lord to whom they owed service and obedience. Already they were tied to the lands on which they laboured and were dependent on their protectors for justice. For many reasons ordinary freemen tended generally to lose their independence and had to seek aid from their more fortunate and powerful neighbours; thus, they lost their standing in the assemblies of their tribe. Everywhere, except in Friesland and parts of Saxony, the nobles wedged themselves between king or duke and the rank and file. They alone could become prelates of the church, and they alone could compete for the possession and enjoyment of governmental rights. At the level below the dukes, the bulk of administrative authority, jurisdiction, and command in war lay with the margraves and counts, whose hold on their charges developed gradually into hereditary right. The commended men and the half free disappeared from the important functions of public life. In the local assemblies they came only to pay dues and to receive orders, justice, and penalties. Their political role was passive. Those lords whose protection was most worth having also had the largest throng of dependents and thus became more formidable to their enemies and to the remaining freemen. Lordship and submission to it were hereditary, and thus the horizon of the dependent classes narrowed until eventually the lord and his officials filled the place of all secular authority and power in their lives. Military strength, the possession of arms and horses, and tactical training in their use were decisive. Most dependent men were disarmed; that became part of their degradation.

Development of serfdom

The accession of the Saxons. Conrad I was quite unequal to the situation in Germany. According to the beliefs of contemporaries, his failure meant that his house was luckless and lacked the prosperity-bringing virtues that belonged to true kingship. On his deathbed in 918, he therefore proposed that the crown, which in 911 had remained with the Franks, should now pass to the leading man in Saxony, the Liudolfing Henry (later called the Fowler). Henry I was elected by the Saxons and Franks at Fritzlar, their ancient meeting place, in 919. With a monarch of their own race, the Saxons now took over the burden and the rewards of being the kingmaking people. The centre of gravity shifted to eastern Saxony, where the Liudolfing lands lay.

The election of Henry I the Fowler

The transition of the crown from the Franks to the Saxons for a time enhanced the self-sufficiency of the south German tribes. The Swabians had kept away from the Fritzlar election. The Bavarians believed that they had a better right to the Carolingian inheritance than the Saxons (who had been remote outsiders in the 9th century) and in 919 elected their own duke Arnulf as king. They, too, wanted to be the royal and kingmaking people. Henry I's regime rested in the main on his own position and family demesne in Saxony and on certain ancient royal seats in Franconia. His kingship was purely military. He hoped to gather authority by waging successful frontier wars and to gain recognition in the first place by concessions rather than to insist on the sacred and priestlike status of the royal office that the church had built up in the 9th century. At his election he refused to be anointed and consecrated by the archbishop of Mainz. In settling with

the Bavarians, he abandoned the policy of supporting the internal opposition that the clergy offered to Duke Arnulf, a plank to which Conrad had clung. To end Arnulf's rival kingship, Henry formally surrendered to him the most characteristic privilege and honour of the crown: the right to dispose of the region's bishoprics and abbeys. Arnulf's homage and friendship entailed no positive obligations toward Henry, and the Bavarian duke pursued his own tribal interests—peace with the Hungarians and expansion across the Alps—as long as he lived.

From these unpromising beginnings the Saxon dynasty not only found its way back to Carolingian traditions of government but soon got far better terms in its relations with the autonomous powers of the duchies, which had gained such a start on it. Nonetheless, the constitution that it bequeathed to its Salian successors was self-contradictory; while seeking to overcome the princely aristocracies of the stem lands by leaving them to themselves, the Saxon kings came to rely more and more, both for the inspiration and for the practice of government, on the prelates of the church, who were themselves recruited from the ranks of the same great families. They loaded bishoprics and abbeys with endowments and privileges and thus gradually turned the bishops and abbots into princes with interests not unlike those of their lay kinsmen. These weaknesses, however, lay concealed behind the personal ascendancy of an exceptionally tough and commanding set of rulers up to the middle of the 11th century. Thereafter, the ambiguous system could not take the strain of the changes fermenting within German society and even less the attack on its values that came from without—from the reformed papacy.

Reliance on the clergy

The Liudolfing kings won military success, and with it they gained that respect for their personal authority that counted for so much at a time when the great followed only those whose star they trusted and who could reward services with the spoils of victory. In 925 Henry I brought Lotharingia back to the East Frankish connection. Whoever had authority in that half-French-speaking, half-German-speaking region could treat the neighbouring kingdom of the West Franks as a dependent. The young Saxon dynasty thus won for itself and its successors a hegemony over the west and the southwest that lasted at least until the mid-11th century. The Carolingian kings of France, as well as the great feudatories who sought to dominate if not to ruin them, became, in turn, petitioners of the German court during the reign of the Ottos. The kings of Burgundy—whose suzerainty lay over the valleys of the Saône and the Rhône, the western Alps, and Provence—fell under the virtual tutelage of the masters of Lotharingia. Rich in ancient towns, this region, once the homeland of the Carolingians, was more thickly populated and wealthier than the lands east of the Rhine. Lotharingian merchants controlled the slave trade from the Saxon marches to Córdoba.

The eastern policy of the Saxons. Greater prestige still and a claim to imperial hegemony fell to the Saxon rulers when they broke the impetus of the Hungarian invasions, against which the military resources and methods of western European society had almost wholly failed for several decades. In 933, after long preparations, Henry routed a Hungarian attack on Saxony and Thuringia. In 955 Otto I (reigned 936–973), at the head of a force to which nearly all the tribes had sent mounted contingents, annihilated a great Hungarian army on the Lech River near Augsburg. The battle again vindicated the efficiency of the heavily armed man skilled in fighting on horseback.

Defeat of the Hungarians at the Lech

With a Saxon dynasty on the throne, Saxon nobles gained office and power, with opportunities for conquest along the eastern river frontiers and marches of their homeland. Otto I indeed had an eastern policy that aimed at getting more than slaves, loot, and tribute. Between 955 and 972 he founded and richly endowed an archbishopric at Magdeburg, which he intended to be the metropolis of a large missionary province among the heathen Slavs beyond the Elbe. This would have brought their tribes under German control and exploitation in the long run; but the ruthless methods of the Saxon lay lords clashed with the church's efforts at more peaceful penetration.

In the 10th century there was little or no German agricultural settlement beyond the Elbe. Far too much forest clearing remained to be done in all the regions of western and southern Germany. The Saxon conquests up to the Oder were secured by military strongholds, called burgwards, and were held only as long as their garrisons had the upper hand. Beyond the Slav peoples of Brandenburg and Lusatia, moreover, new Slavic powers rose: the Poles under Mieszko I and, to the south, the Czechs under the Přemyslids received missionaries from Passau and Magdeburg without falling permanently under the political and ecclesiastical domination of Bavarians and Saxons. The heathen Elbe Slavs, subjugated by the Saxon margraves, rose in 983 when the military occupation collapsed along with the missionary bishoprics that had been founded at Oldenburg, Brandenburg, and Havelberg. Farther south the defenses of the Thuringian marches between the Saale and the middle Elbe remained in German hands, but only after a long and fierce struggle against Polish invaders early in the 11th century. The northern part of the frontier reverted to what it had been before Otto's trustees, Hermann Billung and Gero, opened their wars. Missionary enterprises directed from Bremen and Magdeburg achieved little before the 12th century. The Saxon ruling class, bishops, and margraves must bear the responsibility for the fiasco of eastward expansion in the 10th century. The prelates, too, saw their missions as means to found ecclesiastical empires with subject dioceses and tithes on Slav soil. The tribes across the Elbe therefore remained unconverted and implacable foes, a standing menace to the nearby churches. The wars also left a legacy of savagery on both sides so that from about 1140 onward the substitution of German settlers for the native Slavs became the common policy of both the church and the princes.

Legacy of savagery

Dukes, counts, and advocates. Conrad I's and Henry I's kingships rested on the will of the tribes—or rather on that of their leaders and of the higher aristocracy. It was in the first place an arrangement between the Franks and the Saxons that the Bavarian and Swabian dukes recognized at a price by acts of personal homage, but the German kings, of whatever dynasty, had to live under Frankish law. After the death of Conrad I's brother Eberhard in 939, Otto I kept the Franconian dukedom vacant and the Franconian counts henceforth stood under the immediate authority of the crown. In Saxony, too, Otto kept in his hands the dukedom of his ancestors. The march-duchy of the Billungs, a bulwark raised against the Danes and the northern Slav tribes, did not give the Billung family authority over all the other Saxon princes.

In the south the Ottonians sought to turn the dukedoms of the stem lands into royal fiefdoms and to supplant native dynasties by aliens and members of their own clan. When even that policy did not stop rebellions under the banner of tribal self-interest, they began to break up the ancient Bavarian stem land by creating a duchy in Carinthia to cut off the spearhead of Bavarian expansion southward. The first two Salians, Conrad II (reigned 1024–39) and Henry III (reigned alone 1039–56), also bestowed vacant duchies quite freely on their own kin and on men from outside the stem boundaries. They competed against ducal power but could neither abolish nor replace it. In the 11th century as before, the dukes held assemblies of their folk, led the tribal host in war, and enforced peace.

The counts, who were the ordinary officers of justice in serious, criminal cases, obeyed the ducal summons; but, for the most part, they received their "ban," the power to do blood justice, from the king himself. The fiefs and the customary rights attached to their office, and indeed the office itself, not only became hereditary but also came to be treated more and more as a patrimony to which they had an inherent right against all men, king and duke included. Even so, however, a good many lines died out and their counties fell back into the king's hands. From Otto III's reign (983–1002) onward, it became not at all unusual to bestow these counties on bishoprics and certain great abbeys rather than to grant them out again to other lay magnates. The bishops, however, could not perform all the functions of the counts; in particular, their holy orders forbade them to pass judgments of blood. They often

The rise of the advocates

subinfeudated their countships, and they needed officials called advocates (*Vögte;* singular *Vogt*) to take charge of the higher jurisdiction in the franchises that their churches possessed by royal grant. In the 10th and 11th centuries these advocates had to be recruited from the aristocracy, the very class whose greed for hereditary office was to be checked, because ordinary freemen could not enforce severe sentences or defend the privileges of the church against armed intrusion. Dangerous neighbours of bishoprics and abbeys in any case, the nobles as advocates and protectors of ecclesiastical possessions were anything but reliable servants of their ecclesiastical overlords.

Thus, there arose in nearly all German lands, whether the ducal office survived or not, powerful lines of margraves, counts, and hereditary advocates who enriched themselves at the expense of the church (which meant also the crown) and in competition with one another. From the abler, more fortunate, and long-lived races among these dynasts sprang the territorial princes of the later 12th and 13th centuries, absorbing and finally inheriting most of the rights of government.

Power of the kings

The king was the personal overlord of all the great. His court was the seat of government, and it went with him on his long journeys. The German kings, even more than other medieval rulers, could only make their authority respected in the far-flung regions of their kingdom by traveling ceaselessly from duchy to duchy, from frontier to frontier. Wherever they stayed, their jurisdiction superseded the standing power of dukes, counts, and advocates, and they could collect the profits of local justice and wield some control over it. As they came into each region, they summoned its leaders to attend their solemn crown wearings, deliberated with them on the affairs of the Reich and the locality, presided over pleas, granted privileges, and made war against peacebreakers at home and on enemies abroad.

The promotion of the German church. The royal revenues came from the king's demesne lands and from his share of the tributes that Poles, Czechs, heathen Slavs, and Danes paid whenever he could enforce his claims of overlordship. There were also profits from tolls and mints that had not yet been granted away. The king's demesne was his working capital. He and his household lived on its produce during their wanderings through the Reich, and it also served to provide for his family, to found churches, and to reward faithful services done to him, especially in war. To swell the hosts, vassals had to be enfeoffed, and alienations were inevitable. The Salians, though they inherited the remains of Ottonian wealth as imperial demesne, brought little of their own to make up for its diminution. The last Saxon, Henry II (1002–24), and after him Conrad II, accordingly took to enfeoffing vassals with lands taken from the monasteries. Since the beneficiaries were often already powerful and wealthy men in their own right, no class of freeborn, mounted warriors linked permanently with the crown sprang from the loyalties and rewards of one or two reigns. In any case, the lion's share of grants went to the German church.

The growth of central government

From the Carolingians, the German kings inherited their one and only institution of central government: the royal chapel, with the chancery that does not seem to have been distinct from it. Service there became a recognized avenue of promotion to the episcopate for highborn clerks. In the 11th century, bishops and abbots conducted the affairs of the Reich much more than the lay lords, even in war. They were its habitual diplomats and ambassadors. Unlike Henry I, Otto I and his successors sought to free the prelates from all forms of subjection to the dukes. The king appointed most of them, and to him alone, as to one sent by God, they owed obedience.

Thus, there arose beside the loose association of stem lands in the German kingdom a more compact and uniform body with a far greater vested interest in the Reich: the German church. By ancient Germanic custom, moreover, the founder of a church did not lose his estate in the endowment that he had made; he remained its proprietor and protecting lord. The bishoprics, it is true, and certain ancient abbeys, such as Sankt Gallen, Reichenau, Fulda, and Hersfeld, did not belong to the king; they were members of the kingdom but under his guardianship. The greater churches therefore had to serve the rulers with mounted men, money, and free quarters. Gifts of royal demesne to found or to enrich bishoprics and convents were not really alienations but pious reinvestments, as long as the crown controlled the appointments of bishops and abbots. The church did not merely receive grants of land, often waste, to settle, develop, and make profitable; it was also given, as has been shown, powers of jurisdiction over its dependents. Nor did the kings stint the prelates in other regalian rights, such as mints, markets, and tolls. These grants broke up counties and to some extent even duchies, and that was their purpose: to disrupt the secular lords' jurisdictions that had escaped royal control.

This policy of fastening the church, a universal institution, into the Reich, with its well-defined frontiers, is usually associated with the name of Otto I, but it gathered momentum only in the reigns of his successors. The policy reached a climax under Henry II, the founder of the see of Bamberg in the upper Main valley; nonetheless, Conrad II, though less generous with his grants, and his son Henry III continued it. Bishops and abbots became the competitors of lay princes in the formation of territories, a rivalry that more than any other was the fuel and substance of the ceaseless feuds, the smoldering internal wars in all the regions of Germany for many centuries. The welter and the confused mosaic of the political map of Germany until 1803 is the not-so-remote outcome of these 10th- and 11th-century grants and of the incompatible ambitions that they aroused.

The Ottonian conquest of Italy and the imperial crown. Otto I's marriage with Adelaide (Adelheid), daughter of Rudolph II of Burgundy, and the Italian rivalries between his son Liudolf, duke of Swabia, and Otto's brother Henry I, duke of Bavaria, drew him southward. After 951, expeditions into Italy were a matter for the whole Reich under the leadership of its ruler and no longer just an outlet for the expansion of the south German tribes. For the Saxon military class, too, the south was more tempting than the primeval forests and swamps beyond the Elbe. With superior forces at their back, the German kings gained possession of the Lombard kingdom in Italy. There, too, their overlordship in the 10th and the 11th centuries came to rest on the bishoprics and a handful of great abbeys.

After his victory over the Magyars in 955, Otto I's hegemony in the west was indisputable. By the standards of one chronicler, the Saxon Widukind, he had already become emperor because he had subjected other peoples and enjoyed authority in more than one kingdom. But the right to confer the imperial crown, to raise a king to the higher rank of emperor, had fallen to the papacy, which had crowned Charlemagne and most of his successors. The Carolingian order in the west was still the model and something like a political ideal for all Western ruling families in the 10th century. Otto had measured himself against the political tasks that had faced his East Frankish predecessors and more or less mastered them. To be like Charlemagne, therefore, and to clothe his newly won position in a traditional and time-honoured dignity, he accepted the imperial crown and anointment from Pope John XII in Rome in 962. The substance of his empire was military power and success in war; but Christian and Roman ideas were woven round the Saxon's throne by the writers of his own and the next generation. Although the German kings as emperors did not give the law to the Roman church in matters of doctrine and ritual, they became its political masters for nearly a century. The imperial crown enhanced their standing even among the nobles and knights who followed them to Italy and can hardly have understood or wanted all its outlandish associations. Not only the king but also the German bishops and lay lords thus entered into a permanent connection with an empire won on the way to Rome and bestowed by the papacy.

The formation of the Holy Roman Empire

Otto II (reigned alone 973–983) and above all Otto III (983–1002) were strongly drawn toward their new Mediterranean sphere of action, but Henry II (1002–24) returned to a sober regime centred on Germany and contented himself with three brief Italian expeditions.

The Salians, the papacy, and the princes, 1024–1125. Under Conrad II (reigned 1024–39), the first member of the Rhine-Frankish house known as the Salians, the kingdom of Burgundy fell finally under the overlordship of the German crown, and this tough and formidable emperor also renewed German authority in Italy. His son and successor, Henry III (reigned 1039–56), treated the empire as a mission that imposed on him the tasks of reforming the papacy and of preaching peace to his lay vassals. Without possessing any very significant new resources of power, he gave to his authority an exalted and strained theocratic complexion. Yet, under him, the last German ruler to maintain his hegemony in western Europe, the popes themselves seemed to become mere imperial bishops. He deposed three of them, and four Germans held the Holy See at his command; but lay opposition to the emperor in Germany and criticism of his regime over the church were on the increase during the last years of his reign.

The papal reforms and the German church. More than any other feudal society in early medieval Europe, Germany was divided and torn by the revolutionary ideas and measures of the reformed papacy. Beginning with the pontificate of Leo IX (1048–54)—one of Henry III's nominees—the most determined and inspired spokesmen of ecclesiastical reform placed themselves at the service of the Holy See. Only a few years after Henry III's death (1056), they agitated against lay authority in the church, founded on proprietary rights. They regarded the laity as passive partakers of the sacraments and denied the supernatural status of kingship. Priests, including bishops and abbots, who accepted their dignities from lay lords and emperors at a price, according to the reformers, committed a sin; for these earthly powers could not rightly confer churches at all, nor could they own them. They believed, moreover, that thorough reforms could be brought about only by the exaltation of the papacy so that it commanded the obedience of all provincial metropolitans and was out of the emperor's and the local aristocracy's reach.

Effects of the reformed papacy

The endless repetition of the reformers' teachings in brilliant pamphlets and at clerical synods spread agitation in Italy, Burgundy, and Lotharingia—all parts of the empire. Their new program committed the leaders of the movement to a struggle for power because it struck at the very roots of the regime to which the German church had grown accustomed and on which the German kings relied. The vast wealth that Henry IV's predecessors had showered on the bishoprics and abbeys would, if the new teaching prevailed, escape his control and remain at the free disposal of prelates whom he no longer appointed. Under Roman authority the churches were to be freed from most of the burdens of royal protection without losing any of its benefits. The most fiery spirits in Rome did not flinch from the consequences of their convictions. Their leader Hildebrand, later Pope Gregory VII (reigned 1073–85), was ready to risk a collision with the empire.

Henry IV was not yet six years old when his father died in 1056. The full impact of the Gregorian demands—coming shortly after a royal minority, a Saxon rising, and a conspiracy of the south German princes—has often been regarded as the most disastrous moment in Germany's history during the Middle Ages. In fact, the German church proved thoroughly unreliable as an inner bastion of the empire even before Rome struck. Its leaders, Anno and Adalbert, archbishops of Cologne and of Hamburg–Bremen respectively, shamelessly exploited their hold over the young king by hunting for spoils out of the imperial demesne. In 1074 and 1075 Gregory proceeded against simony (the buying and selling of church office) in Germany and humiliated the aristocratic episcopate by summonses to Rome and sentences of suspension. These papal actions demoralized and shook the German hierarchy. The prelates' return to their customary support of the crown was not disinterested, nor wholehearted, nor unanimous.

Decrees of Gregory VII

The discontent of the lay princes. Henry IV's minority also gave elbowroom to the ambitions and hatreds of the lay magnates. The feeble regency of his mother, Agnes of Poitou, faltered before the throng of princes, who respected only authority and forces greater than their own. The ruling influence of the higher clergy at the court of Henry III and the renewed flow of grants to the church had estranged them from the empire. It is likely also that these eternally belligerent men were lagging behind the prelates in the development of their agrarian resources. The prelates had a vested interest in peace, and under royal protection they improved and enlarged their estates by turning forests into arable land and also by offering better terms to freemen in search of a lord. The bishops' market and toll privileges brought them revenues in money, which many of the lay princes lacked. So far, however, the princes' military power, their chief asset, had remained unchallenged. Now, for the first time, they also had to face rivals within their own sphere of action. Henry III and the young Henry IV began to rely on advisers and fighting men drawn from a lower tier of the social order—the poorer, freeborn nobility of Swabia and, above all, the class of unfree knights, known as *ministeriales.* These knights had first become important as administrators and soldiers on the estates of the church early in the 11th century. Their status and that of their fiefs was fixed by seignorial ordinances, and they could be relied on and ordered about, unlike the free vassals of bishops and abbots. Beginning with Conrad II, the Salian kings used *ministeriales* to administer their demesne, as household officers at court and as garrisons for their castles. They formed a small army, which the crown could mobilize without having to appeal to the lay princes, whose ill will and antipathy toward the government of the Reich grew apace with their exclusion from it.

Having come of age, Henry IV used petty south German nobles and his *ministeriales* to recover some of the crown lands and rights, which the lay princes and certain prelates had acquired during his minority, particularly in Saxony. His recovery operations went further, however, and a great belt of lands from the northern slopes of the Harz Mountains to the Thuringian Forest was secured and fortified under the supervision of his knights to form a royal territory, where the king and his court could reside. The south German magnates were thus kept at a distance when Henry and his advisers struck at such neighbouring Saxon princes as Otto of Northeim and the Billung family.

Formation of a royal territory

The storm broke in 1073. A group of Saxon nobles and prelates and the free peasantry of Eastphalia, who had to bear the brunt of statute labour in the building of the royal strongholds, revolted against the regime of Henry's Frankish and Swabian officials. To overcome this startling combination and to save his fortresses, the king needed the military strength of the south German princes Rudolf of Rheinfelden, duke of Swabia; Welf IV, duke (as Welf I) of Bavaria; and Berthold of Zähringen, duke of Carinthia. Suspicious and hostile at heart, they took the field for him only when the Eastphalian peasantry committed outrages that shocked aristocratic caste feeling everywhere. Their forces enabled Henry to defeat the Saxon tribal rebellion at Homburg near Langensalza in June 1075. But, when the life-and-death struggle with Rome opened only half a year later, the south German malcontents deserted Henry and, together with the Saxons and a handful of bishops, entered into an alliance with Gregory VII. Few of them at this time were converted to papal reform doctrines, but Gregory's daring measures against the king gave them a chance to come to terms with one another and to justify a general revolt.

The civil war against Henry IV. On Feb. 22, 1076, the pope absolved all men from their oaths to Henry and solemnly excommunicated him. In October Gregory's legates met the German lords at Tribur (modern Trebur) to decide on the future of the king, whom his last adherents now abandoned. Although Henry was absolved by Gregory at Canossa in January 1077, the princes two months later elected Rudolf of Rheinfelden to rule in his place.

The war that now broke out lasted for almost 20 years. A majority of the bishops, most of Rhenish Franconia (the Salian homeland), and some important Bavarian and Swabian vassals sided with Henry. He thus held a central position, dividing his south German from his Saxon enemies, who could not unite long enough to destroy him. With the death in battle of Rudolf of Rheinfelden (1080) and the demise of another antiking, Hermann of

Division of Henry's enemies

Salm (1088), the war in Germany degenerated into a number of local conflicts for the possession of bishoprics and abbeys. It almost died down in 1098, when the south German adherents of the papacy came to terms with Henry for the time being, though without recognizing his antipope Clement III. Throughout these years the crown, the churches, and the lay lords had to enfeoff more and more *ministeriales* in order to raise mounted warriors for their forces. Though this recruitment and frequent devastations strained the fortunes of many nobles, they knew how to recoup themselves by extorting more fiefs out of neighbouring bishoprics and abbeys. The divided German church thus bore the brunt of the costs of civil war, and it needed peace almost at any price.

Henry V and results of the conflict. The Salian dynasty and the rights for which it fought were saved because Henry IV's son and heir himself seized the leadership of a last rising against his father (1105). This maneuver enabled Henry V (reigned 1106–25) to continue the struggle for the crown's prerogative over the empire's churches against the inexorable demands of the papacy. The conflict now shrank into a legalistic dispute over the right to invest bishops and abbots with their dignities and the secular possessions attached to them. As the struggle continued, the princes became the arbiters and held the balance between their overlord and the pope. In 1122, acting as intermediaries and on behalf of the Reich, they forced the temporary concessions known as the Concordat of Worms out of the Holy See and its German spokesman, Archbishop Adalbert of Mainz, the bitter personal enemy of Henry V and the territorial rival of the Hohenstaufen sons of Henry's sister Agnes. By then, however, the princes had for the most part defeated efforts to restore royal rights in Saxony and to stem the swollen jurisdictions and territorial powers of the aristocracy elsewhere.

German loss of power after the death of Henry V

When Henry V, the last Salian, died childless in 1125, Germany was no longer the most effective political force in Europe. The brilliant conquest states of the Normans in England and Sicily and the patient, step-by-step labours of the French kings were achieving forms of government and concentrations of military and economic strength that the older and larger empire lacked. The papacy had dimmed the empire's prestige, and Rome became the true home of universalistic causes. When Pope Urban II preached the first crusade in 1095, Henry IV, cut off and surrounded by enemies, was living obscurely in a corner of northern Italy. The Holy See, by its great appeal to the militant lay nobility of western Europe, thus won the initiative over the empire. At this critical moment the Reich also lost control in the Italian bishoprics and towns, just when their population, trade, and industrial production were expanding fast. Germany did not even benefit indirectly from the crusaders' triumphs, although some of their leaders (*e.g.*, Godfrey of Bouillon and Robert II of Flanders) were vassals of the emperor. The civil wars renewed for a time the relative isolation of the central German regions.

Internally, the crown had saved something of the indispensable means of government in the control over the church; but it was a bare minimum, and its future was problematic. The ecclesiastical princes henceforth held only their temporal lands as imperial fiefs, for which they owed personal and material services. As feudatories of the empire, they came to represent the same interests toward it as did the lay princes; at least, their sense of a special obligation tended to weaken. The king's jurisdiction continued to exist alongside and in competition with that of the local powers. The great tribal duchies survived as areas of separate customary law. Each developed differently, and the crown could not impose its rights on all alike or change the existing social order. The most tenacious defenders of this legal autonomy had been the Saxons; but it also prevailed in Swabia, where distinct territorial lordships grew fast.

The Gregorian reform movement therefore aggravated the age-old contradictions in Germany's early medieval constitution, but its monastic culture and its intellectual interests were anything but barren. Both sides fought with new literary weapons to work on public opinion in cathedrals and cloisters and perhaps also in the castles of the lay aristocracy. In their hard-hitting polemical writings they attempted to expound the fundamental theological, historical, and legal truths of their cause. The agitation did something to disturb the cultural self-sufficiency of the German laity. It drove many of the south German nobles to maintain direct connections with the Holy See, and, whether they wanted to or not, they had to fall in with the aspirations of the religious leaders. The reform movement of the 11th and 12th centuries, it might almost be said, very nearly completed the conversion of Germany that had begun five centuries before.

GERMANY AND THE HOHENSTAUFEN, 1125–1250

Dynastic competition, 1125–52. The nearest kinsmen of Henry V were his Hohenstaufen nephews—Frederick, duke of Swabia (1105–47), and his younger brother Conrad, the sons of Henry's sister Agnes and Frederick, the first Hohenstaufen duke of Swabia. Some form of election had always been necessary to succeed to the crown, but, before the great civil war, nearness to the royal blood had been honoured whenever a dynasty failed in the direct line. By 1125, however, the princes, guided by Archbishop Adalbert of Mainz, no longer respected blood right. Affinity with Henry V was no recommendation to them, and hereditary succession seemed to lower their authority in the government of the Reich. Instead of Frederick they chose the duke of Saxony, Lothair of Supplingenburg (reigned as king 1125–37, reigned as emperor 1133–37). Like the Hohenstaufen, he had risen by a lucky marriage and a successful career of continuous fighting into the first rank of dynasts; but, unlike them, he had served the cause of the Saxon opposition to the Salians.

Disregard for blood right

With the enormous Northeim and Brunonian inheritances behind him, Lothair III (sometimes called Lothair II) could humble the Hohenstaufen brothers (1134) after marrying his only daughter and heiress to a Welf, Henry the Proud. Even without this dazzling alliance, the Welfs, already dukes of Bavaria and possessors of vast demesnes, countships, and ecclesiastical advocacies there, in Saxony, and in Swabia, were somewhat better off than their Hohenstaufen rivals. On the death of Lothair in 1137, however, the fears of the church and a few princes turned against the Welfs. Instead of Henry the Proud, who now held the duchies of Saxony and Bavaria and the Mathildine lands in Italy, they chose Conrad (reigned 1138–52), who had been Lothair's unsuccessful Hohenstaufen opponent.

The battle against the Welfs, which Conrad III put foremost on his political program, was abandoned with his death in 1152, when an election once again decided the succession and the political situation in Germany for the next 30 years. The princes then chose Frederick I Barbarossa (reigned as king 1152–90; as emperor 1155–90), the son of Conrad's elder brother Frederick and the Welf princess Judith. Frederick I agreed to share power in Germany with his Welf cousin Henry the Lion. The price of his election was dualism. In 1156 the duchy of Bavaria, which Conrad had tried to wrest from the Welfs, was restored to Henry the Lion, already undisputed duke of Saxony. The Babenberg margrave of Austria, Henry's rival, had to be compensated with a charter that raised his margravate into a duchy and gave him judicial suzerainty over an even wider area. Taken out of the Lion's duchy, it was to be held as an imperial fief that might descend both to sons and daughters. A perpetual principality, it served as a model for the aspirations of many other lay princes.

Colonization of the east. The history of Germany in the 12th and 13th centuries is one of ceaseless expansion. A conquering and colonizing movement burst across the river frontiers into the swamps and forests from Holstein to Silesia and overwhelmed the Slav tribes between the Elbe and the Oder. Every force in German society took part: the princes, the prelates, new religious orders, knights, townsmen, and peasant settlers. Agrarian conditions in the older lands of Germanic occupation seem to have favoured large-scale emigration. With a rising population, there was much experience in drainage and wood clearing but a diminishing fund of spare land to be attacked in the west. Excessive subdivision of holdings impoverished tenants and did not suit the interests of their lords. Some-

The conquering and colonizing movement of the 12th and 13th centuries

times also, seignorial oppression is said to have driven peasants to desert their masters' estates. They certainly found a better return for their labour in the colonial area: personal freedom, secure and hereditary leasehold tenures at moderate rents, and, in many places, quittance from services and the jurisdiction of the seignorial advocate.

The colonists brought with them a disciplined routine of husbandry, an efficient plow, and orderly methods in siting and laying out their villages. Very soon, even the Slav rulers of Bohemia and Silesia were competing for immigrants. First and foremost, however, the princes of the Saxon and Thuringian marches sought to attract settlers for the lands that they had conquered and the towns that they had founded to open up communications and trade routes. The older regions of the Reich, moreover, had not only peasants but also men of the knightly class to spare—soldiers who needed fiefs and lordships to uphold their rank. Both could be gained beyond the Elbe under the leadership of successful princes. The Germanized east thus became the home of fair-sized principalities in the 13th century, while all along the Rhine River valley the rights of government were tending to be scattered over smaller and less compact territories. The Ascanian dynasty, for instance, which under Albert the Bear began to advance into Brandenburg, by 1250 not only ruled over a broad belt of land up to the Oder River but had already established itself on the eastern banks ready for further advances. Farther south the Wettin margraves of Meissen busied themselves with settlements and town foundations in Lusatia.

For a time Henry the Lion, as duke of Saxony (1142–80), overshadowed all these rising powers, and the Welf profited as much by the ruthless use of his resources against weaker competitors as by his own efforts in Mecklenburg. As his was the only protection worth having in northeastern Germany, the newly established Baltic bishoprics were at his mercy, and he alone could attract the traders of Gotland to frequent the young port town of Lübeck, which he extorted from one of his vassals in 1158.

The Reich, too, possessed demesnes in the east, notably the Egerland, Vogtland, and Pleissnerland in the Thuringian March. The Hohenstaufen kings therefore took some part in opening up these regions. They, too, founded towns and monasteries on their thickly wooded lands and established their *ministeriales* as burgraves and advocates over them. But in this, as in many other things, they only competed with the princes. They did not and could not control the eastward movement as a whole.

Hohenstaufen policy in Italy. In the other great field of German expansion in the 12th century—Lombardy and central Italy—the emperors and their military following

Adapted from J.R. Strayer and D.C. Munro, *The Middle Ages, 395–1500*, 5th ed., p. 521, copyright © 1970, by permission of Appleton-Century-Crofts, Educational Division, Meredith Corporation



German expansion to the east, 800–1400.

alone counted. The rural population of Germany had no direct interest in the wars waged to recover and exploit regalian rights over the growing Lombard city communes. The connection between the German crown, the empire, and dominion over Italy has indeed been regarded as a disaster for Germany, and the ever-increasing concern of the Hohenstaufen dynasty with the south as its most tragic phase. Although Frederick Barbarossa's policy was opportunistic, he had really very little choice. Having bought off the Welfs and reconciled other great families with yet more concessions and lastly endowed his own cousin, Conrad III's son Frederick, with Hohenstaufen demesnes in Swabia, he had to try to mobilize their goodwill for the empire while it lasted. He now aimed to set up a regime of imperial officials and captains who were to exact dues and to control jurisdiction that the communes had usurped from the failing grasp of their bishops. The Germans in Italy did not bring valuable accomplishments to poor and primitive tribesmen, but they attacked economically advanced and better developed communities, to which they had nothing to offer in return for the rights and taxes they demanded. Military power was their chief asset in Lombardy, and they used it ruthlessly.

The Italian policy

For the Hohenstaufen *ministeriales* the rule of their masters in northern and central Italy was a career. Because they could be deployed continuously, they became the backbone of the imperial occupation. A handful of minor dynasts also served Barbarossa for many years in the powerful and profitable commands that he established. The German bishops and certain abbots still had to supply men and money, and some of them threw themselves wholeheartedly into the war: for instance, Rainald of Dassel and Philip of Heinsberg, archbishops of Cologne from 1159 to 1167 and from 1167 to 1191 respectively, who, as archchancellors for Italy, had a vested interest in it. The support of the lay princes, conversely, was fitful and sporadic. Even at critical moments they could not be counted on unless they individually agreed to serve or to send their much-needed contingents for a season. The refusal of the greatest of them, Henry the Lion, in 1176 brought about the emperor's defeat at the Battle of Legnano and spoiled many years' efforts in Lombardy.

The fall of Henry the Lion and the estate of princes. Forced to retreat before the papacy and the Lombard League in 1177, Barbarossa cooled toward his Welf cousin, whom he could justly blame for some of his setbacks. Dualism in Germany had outlived its purpose. Hitherto, the enemies of Henry—the princes, bishops, and magnates of Saxony—had been unable to gain a hearing against him at the emperor's court days. By 1178, however, the emperor was ready to help them. Outlawed (1180), beaten in the field and deserted by his vassals, Henry had to surrender and go into exile in 1182. His duchies and fiefs were forfeited to the Reich.

Surrender and exile of Henry the Lion

His fall left a throng of middling princes face to face with an emperor whose prestige, despite reverses, stood high and whose resources had greatly increased since he began to reign. The princes were nonetheless the chief and ultimate gainers from the events of 1180. The final judgment by which Henry the Lion lost his honours was not founded on folk law but on feudal custom. The princes who condemned him regarded themselves as the first feudatories of the empire, and they decided on the redistribution of his possessions among themselves. During the 12th century the stem duchies of the Ottonian period finally disintegrated. Within their ancient boundaries not only bishops but also lay lords succeeded in eluding the authority of the dukes. In their large immunities they themselves wielded stem-ducal powers. To enforce the imperial peace laws became both their ambition and their justification. Everywhere the greater lay dynasties and even some bishops tried to acquire a ducal or an equivalent title that would enable them to consolidate their scattered jurisdictions and, if possible, to force lesser freelords to attend their pleas.

These highest dynasts had interests in common, and they closed their ranks not only against threats from above but also against fellow nobles who had been less successful in amassing wealth, counties, and advocacies and who did not possess the superior jurisdiction of a duke, a margrave, a count palatine, or a landgrave. They and they alone were now called princes of the empire. To lend a certain cohesion to their varied rights, they were willing to surrender their houselands to the Reich and receive them back again as a princely fief. For the emperor it was theoretically an advantage that men so powerful in their own right should owe their chief dignity and most valued privileges to his grant. It opened the possibility of escheats (reversions), for in feudal custom the rules of inheritance were stricter than in folk right. In Germany, however, the political misfortunes of rulers succeeded, by and large, in ensuring that ancient caste feeling and notions of inalienable right conquered the principles of feudal law. By 1216 it was established that the emperor could neither abolish principalities nor create princes at random.

The "heirs" of Henry the Lion had to fight a ceaseless battle to establish and maintain themselves. In Bavaria the Wittelsbachs had received the vacant duchy, but they were not recognized as superiors by the dukes of Styria or by the dukes of Andechs-Meran. In Saxony the archbishop of Cologne was enfeoffed with Henry the Lion's ducal office and with all his rights in Westphalia, while an Ascanian prince, Bernard of Anhalt, received the eastern half of Henry's duchy. Neither Bernard nor the archbishop, however, could make much out of their dukedoms, except in the regions where they already had lands and local jurisdictions. All over the Reich these and regalian rights, such as mints, fairs, tolls, and the right of granting safeconducts, were the substance of princely power. To possess them as widely as possible became the first goal of the abler bishops and lay lords.

The Hohenstaufen conflict with the papacy, 1159–1215. The attempt to establish a direct imperial regime in Italy antagonized the papacy once again and led to a new struggle with Rome, the ally of the Lombard communes. Political and territorial rather than ecclesiastical interests were at stake; but the popes could only fight as heads of the universal church, defending its liberty against a race of persecutors, and they had to employ their characteristic weapons—excommunication, propaganda, and intrigue. Nonetheless, the German bishops stood by Barbarossa and, for the most part, followed him in maintaining a prolonged schism against Pope Alexander III. Unsuccessful in Lombardy, the centre of Hohenstaufen ambitions after 1177 shifted to Tuscany, Spoleto, and the Romagna. This redoubled the fears and the resentment of the popes, particularly after 1189 when Frederick's son and chosen successor, Henry VI (reigned 1190–97), became the legitimate claimant to the Sicilian kingdom through his wife Constance, the sole surviving legitimate heiress.

With their backs to the wall, the popes had to make what use they could out of any opposition to the Hohenstaufen. Their chance came in 1197 when Henry VI died prematurely, leaving a three-year-old son, Frederick, to succeed him. To escape the chaos of a minority regime, the bulk of the German princes and bishops in 1198 elected the boy's uncle Philip of Swabia; but an opposition faction in the lower Rhenish region, led by the archbishop of Cologne and financed by Richard I of England, raised an antiking in Otto IV, a younger son of Henry the Lion. Pope Innocent III had to enlarge on his rights over imperial coronations and become a partisan in the German electoral feud if he wished to defend his recovered holdings in Italy against Hohenstaufen claims. Territorial interests in the Romagna tempted the papacy to exploit the weaknesses of the empire's constitution, the uncertainties of electoral custom, and the lack of strict legal norms in Germany. During the war for the crown, much hard-won demesne and useful rights over the church had to be sacrificed by the rivals to bribe their supporters.

Civil war

Frederick II and the princes. Henry's son Frederick II entered Germany to regain his own against Otto IV in 1212 and secured the crown in 1215. Despite promises to divide his inheritance, he kept the kingdom of Sicily and the empire together, and thus he also became locked in the inevitable life-and-death struggle with the papacy. The Hohenstaufen demesne in Swabia, Franconia, and Alsace and on the middle Rhine was still very considerable, and

Frederick even recovered certain fiefs and advocacies that had been lost during the earlier civil wars. Their administration was improved, and they provided valuable forces for his Italian wars. The great peace legislation of 1235, moreover, showed that the emperor had not become a mere competitor in the race for territorial gain. But, except for brief intervals, the princes and bishops were left free to fight for the future of their lands against one another and against the intractable lesser dynasts who refused to accept their domination. The charters that Frederick had to grant to the ecclesiastical princes (the so-called *Confoederatio cum principibus ecclesiasticis,* 1220) and later to all territorial lords (*Constitutio,* or *Statutum in favorem principum,* 1232) gave them written guarantees against the activities of royal demesne officials and limited the development of imperial towns at the expense of episcopal territories. But the charters were not always observed, and until 1250 the crown remained formidable in southern Germany, despite the antikings Henry Raspe and William of Holland, whom the papacy caused to be elected by the Rhenish archbishops in Germany in 1246 and 1247.

The Reich after the Hohenstaufen catastrophe. Frederick II died in 1250, in the midst of his struggle against Pope Innocent IV. His son Conrad IV left the north in 1251 to fight for his father's Italian possessions. William of Holland, antiking from 1247 to 1256, was thus without a rival in an indifferent Germany that had lost interest in its rulers. The bishops' cities and the towns, many of them founded on royal demesne, could not be absorbed. Their economic power challenged the age-old aristocratic order in German society, and, deprived of royal protection, they banded together to defend their autonomy. Within the nobility each rank tended to acquire some of the personal rights of its betters. The Hohenstaufen breakdown after 1250 left a gap in Swabia that no rising territorial power was able to fill. Countless petty lords and imperial *ministeriales* of the southwest succeeded in holding their seigniories as immediate vassals of the Reich. Their independent territories often survived for centuries.

Rise of the *ministeriales*

The *ministeriales* elsewhere, too, ceased to be the dependable servants that they once had been. Many free nobles voluntarily joined their ranks, and the knights thus assimilated the rights of the free aristocracy. They became the governing class of the territorial principalities, the standing councillors of their masters, whose household offices and local justice they monopolized and held in fee for many generations. Without the consent of this territorial nobility, the princes could neither tax nor legislate. Even the less important *ministeriales,* who only administered manors for their lords, entrenched themselves as hereditary bailiffs, who kept surplus produce for themselves and usurped seignorial dues, so that it paid the owners to commute the labour services of their villeins into money rents and to lease out those portions of the demesne that the unfree peasants had cultivated for them. Even then, however, the hereditary officials could not be easily dislodged. Finally, the ambitions of the princes themselves did not aim above the patrimonial policies of the past. They were acquisitive and attempted to build up their territories by usurpation, inheritance, marriage treaties, and escheats. They also tried, where possible, to administer their lands with officials whom they could depose at will. Yet they did so not to found sovereign states but chiefly to provide for their families. Again and again, they divided their dominions among sons, who, in turn, founded cadet lines and set them up on a fraction of the principality.

The loss of effective central authority

By 1250 there was thus no really effective central authority left in Germany. The prince-bishoprics had become fiercely contested prizes between neighbouring dynasties, often vassals of the see (*i.e.,* the bishopric). But constant feuds, disorder, and insecurity did not, by any means, frustrate the immense energies of the Germans in the 13th century. Eastward expansion continued under the leadership of the princes and, above all, of the Knights of the Teutonic Order. Their advance into Prussia went hand in hand with the opening up of the Baltic by the merchants of Lübeck. It is possible that three centuries of complete security from foreign invasion made it unnecessary for the German aristocracy to learn the virtues of political self-discipline and subordination; but it would be a great mistake to judge Hohenstaufen Germany solely by its failure to achieve political and administrative unity. (K.J.L.)

Germany from 1250 to 1493

1250 TO 1378

Decline of Hohenstaufen power

The extinction of the Hohenstaufen dynasty. The death of Frederick II in 1250 and of his son Conrad IV in 1254 heralded the irreversible decline of Hohenstaufen power in Germany and in the conjoint kingdoms of Naples and Sicily. Conrad's infant son Conradin, heir to Naples and Sicily, remained in Germany under the guardianship of his Bavarian mother. His uncle Manfred seized the reins of government in both Italian kingdoms and in 1258 formally supplanted Conradin by engineering his own coronation in Palermo. Manfred's defiance of papal claims to suzerainty over the kingdoms impelled the French-born Pope Urban IV to grant them to Charles of Anjou, brother of Louis IX of France. Papal taxation of the French clergy and loans from Florentine bankers enabled Charles to raise a large mercenary army for an expedition to Italy. Manfred, deserted by his barons, was defeated and slain near Benevento in 1266. Conradin then rallied his German supporters and led them across the Alps. But Conradin's financial resources were inadequate; unpaid troops deserted, and his depleted following was routed by Charles near Tagliacozzo (1268). Conradin was captured as he fled toward Rome, convicted of lese majesty (a form of treason), and beheaded in the public square at Naples.

The Great Interregnum. In Germany, the death of Frederick II ushered in the Great Interregnum (1250–73), a period of internal confusion and political disorder. The antikings Henry Raspe (landgrave of Thuringia, 1246–47) and Count William of Holland (ruled 1247–56) were elected by the leading ecclesiastical princes at the behest of the papacy. William's title was recognized initially only in the lower Rhineland, but his marriage to Elizabeth of Brunswick in 1252 ensured his acceptance by the interrelated princely dynasties of north Germany. The death of the Hohenstaufen Conrad IV left William without a rival in Germany. His growing strength and independence enabled him to escape from the tutelage of his ecclesiastical electors and to devote himself to purely dynastic policies. He pursued his feud with Margaret, countess of Flanders, over their conflicting territorial claims in Zeeland at the mouth of the Rhine. He renewed the attempts of his dynasty to obtain complete mastery of the Zuider Zee by thrusting eastward at the expense of Friesland; he died at the hands of the Frisians in 1256.

Initiative of the German princes

Pope Alexander IV forbade the election of a Hohenstaufen but interfered no further with the succession. Hence the initiative was taken by a small group of influential German princes, lay and ecclesiastical, acting out of self-interest. None desired the election of a ruler powerful enough to threaten their growing independence as territorial princes; nor did they single out a German candidate, who might prove to be as uncontrollable as William. Archbishop Conrad of Cologne approached Richard, Earl of Cornwall, brother of Henry III of England. Richard's gifts and assurances of future favour brought him the votes of the archbishops of Cologne and Mainz, the count palatine of the Rhine, and Otakar II of Bohemia. He was formally elected in 1257 and crowned king at Aachen (Aix-la-Chapelle). Three months after Richard's election, Alfonso X of Castile, who aspired to the empire in order to strengthen his foothold in Italy, was chosen in similar fashion by the archbishop of Trier, the duke of Saxony, the margrave of Brandenburg, and the devious Otakar.

The candidates were distracted by the turbulence of the aristocracy in their countries—Richard paid four fleeting visits to Germany; Alfonso failed to appear at all. Both appealed to the papacy for confirmation of their election. Their claims were summarized in Urban IV's bull *Qui Coelum* (1263), which assumed that the exclusive right of election lay with the seven leading princes involved in the double election of 1257.

The rise of the Habsburgs and Luxembourgs. *Rudolf of Habsburg.* When Richard died in 1272 the electoral

princes were spurred into action by Pope Gregory X, who desired the election of a German monarch sympathetic toward a crusade for the recovery of the Holy Land. The princes, dreading an overly powerful king, rejected the advances of Philip III of France and Otakar. They chose instead Rudolf of Habsburg (1273), a minor count of Swabia who lacked the strength to regain the crown domains the electors had usurped during the Interregnum. Papal diplomacy persuaded Alfonso X to abandon his pretensions to the throne; but Otakar denounced the election on the ground that the duke of Bavaria had voted as lay elector in his stead. Rudolf allied himself with the Wittelsbach family of Bavaria and with other envious neighbours of Otakar, who was defeated and slain (1278). The duchies of Austria and Styria, overrun by Otakar during the Interregnum, were declared vacant and conferred jointly on Rudolf's sons Albert and Rudolf (1282). These acquisitions placed the Habsburgs in the first rank of the German territorial princes and lent impetus to a gradual shift in the political centre of gravity from the Rhineland to east Germany. The growing Habsburg power, however, disquieted the electoral princes, who frustrated the king's attempts to secure the election of his elder son Albert in 1287 and of his younger son Rudolf in 1290.

Growing Habsburg power

Adolf of Nassau. On the death of Rudolf in 1291, the electors averted the danger of a hereditary Habsburg monarchy by choosing Count Adolf of Nassau as his successor. Adolf, possessing only a small patrimony to the south of the river Lahn, strengthened himself financially by promising military aid to and receiving subsidies from both sides in the current Anglo-French war. He took possession of Meissen when the cadet branch of the Wettin dynasty died out, and he used his foreign subsidies to purchase Thuringia in 1295. He was thus able to adopt a more independent attitude toward his electors. On June 23, 1298, five of the electors pronounced Adolf unfit to rule and deposed him; on the following day they elected Albert of Austria in his stead. Albert marched westward from Austria at the head of a large army, and in a battle at Göllheim, Adolf was slain and his supporters fled.

Albert I of Habsburg. By restoring the Habsburg Albert I (ruled 1298–1308) to the kingship, the electors placed themselves in jeopardy. The new ruler, backed by the ample resources of his Austrian dominions, was more powerful and unscrupulous than his predecessor. The electors regarded his treaty of friendship with Philip IV of France (1299) as a move to enlist French support for the election of his son Rudolf as his successor in Germany. His attempt to seize Holland and Zeeland as a vacant fief of the empire was rightly interpreted by the electors as an effort to establish Habsburg influence on the lower Rhine (1300). The four prince-electors of the Rhineland (the archbishops of Mainz, Trier, and Cologne and the count palatine) conspired to depose Albert. But Albert wrecked the design by decisive military action (1301–02), and he sealed his victory over the electors by obtaining confirmation in 1303 of his election from Pope Boniface VIII in return for an unprecedented oath of fealty and obedience to the papacy. Albert subsequently renewed Adolf's claims to Meissen and Thuringia, but his authority there was still disputed when he died by assassination in 1308. Albert temporarily tamed the electoral princes, placated the papacy, and renounced intervention in Italy; but this policy foundered at his death, and the electors were given a fresh opportunity to reassert their influence over the German monarchy.

Oath of fealty and obedience to the papacy

Henry VII of Luxembourg. The princes, released from Albert's heavy hand, sought a servant, not a master. Archbishop Baldwin of Trier sponsored the candidature of his brother, Count Henry of Luxembourg, who was elected at Frankfurt (modern Frankfurt am Main) in 1308 as Henry VII. The house of Luxemburg (Luxemburg) was not a major territorial power, and Henry lost no time in exploiting his new status to extend its possessions. Under his direction the Diet of Frankfurt (1310) closed the long-disputed question of the Bohemian succession by awarding the kingdom, with the consent of the Bohemian estates, to Henry's son John. Thus, in common with the Habsburgs, the main weight of Luxembourg interests gravitated eastward. But Henry, unlike his Habsburg predecessors, dreamed of a restoration of the ancient authority of the empire in Italy. His Italian expedition (1310–13) opened brilliantly, and in 1312 he was crowned Holy Roman emperor at Rome. The old fear of German domination, however, stiffened the resistance of the Italian states. Pope Clement V was alarmed by Henry's preparations to invade the kingdom of Naples, a papal fief, and threatened excommunication. A renewed collision of empire and papacy seemed imminent when Henry died in 1313.

The growth of territorialism under the princes. The decline of Hohenstaufen influence in Germany, the Interregnum, and the rapid alternation of dynasties on the German throne created favourable conditions for the territorial princes, lay and spiritual, to gain power. Frederick II had purchased the support of the princes by lavish grants of crown lands, chiefly in the Rhineland and Thuringia; in 1220 he procured the cooperation of the ecclesiastical princes in the election of his son Henry as king and eventual heir to the empire by renouncing his regalian rights of building castles, issuing coinage, and imposing tolls on merchandise in their territories. Henry himself had extended similar concessions to the lay princes in 1231.

Thenceforth the direct action of royal authority was virtually precluded in the princely domains. The princes were at liberty to multiply castles and toll stations, establish mints, exploit mineral deposits, and settle all judicial cases except those transferred on appeal to the court of the emperor. The machinery of administration under the prince and his council (*Hofrat*) was, nevertheless, still rudimentary. Public taxation was intermittent and restricted to emergency occasions, and it was subject to the consent of the three estates of the principality (clergy, nobles, townspeople), which were consulted separately by the prince. The estates grasped the opportunity to ventilate their grievances and to press their advice upon the prince. The emerging territorial state was thus under the dual government of the prince and the estates, and its development was to be heavily influenced by a shifting balance of power between them.

Downfall of royal authority in princely domains

Constitutional conflicts in the 14th century. The death of Henry VII led to a disputed election and a civil war in Germany. The electors' impulse to choose another lesser count as king was checked by the houses of Habsburg and Luxembourg, which pressured the prince-electors to choose between their candidates. The pro-Habsburg majority elected Frederick the Handsome, duke of Austria. The minority withdrew their support from Henry VII's son John and transferred it to a more formidable candidate, the Bavarian duke Louis of Wittelsbach, who had recently broken an Austrian invasion of his duchy.

Electoral custom did not yet acknowledge the majority principle. The papacy, which had claimed the right to adjudicate disputed elections since 1201, was vacant. Hence the two claimants settled their differences by the sword. In 1322 Louis defeated and captured his rival at Mühldorf, but his triumph in Germany merely raised the curtain on a long and bitter dispute with the papacy.

Pope John XXII, guided by canon law and precedent, affirmed that Louis might not legally rule until confirmed by the papacy; thus the disputed election of 1314 and the absence of papal approbation invalidated Louis' royal title and his right to govern. Louis contended, however, that election by a majority conferred a legitimate title and administrative power and did not require papal confirmation. His defiance of the pope exposed him to excommunication (1324) and to the procedures of canon law, whereby he was required to submit entirely to the papal terms before absolution could be granted. Louis warned the electors that their rights were endangered by the subjection of the elections to papal confirmation. Six electors responded in the Declaration of Rhens (1338), proclaiming as an ancient custom of the empire that election by a majority was valid and that the king-elect assumed his administrative power immediately, without the intervention of papal approbation. Under Louis' direction the declaration was repeated at the subsequent Diet of Frankfurt as an imperial law, and offenders against it were declared guilty of lese majesty.

The Declaration of Rhens

John XXII and his successors were unyielding. In 1343 Pope Clement VI made diplomatic overtures to Charles of Luxembourg, heir to the Bohemian throne, with the object of procuring his election to the German kingship in Louis' stead. The electors, led by Baldwin of Luxembourg, archbishop of Trier, began to desert Louis one by one. The pope thereupon urged a new election. Charles assured the pope secretly that he would await papal confirmation of his forthcoming election before exercising governmental power in the Italian possessions of the empire, but, despite intense pressure by Clement, he would accept no such restriction with regard to Germany. In 1346 only two electors remained faithful to Louis: his son Louis of Brandenburg and his kinsman Rudolf, count palatine of the Rhine. The other five assembled at Rhens on July 11 and elected Charles under the title of Charles IV. The new king was spared a lengthy conflict with his rival, who died of apoplexy in 1347.

Charles IV and the Golden Bull. Charles IV (ruled 1346–78) readily perceived that disputed elections exploding into civil war had been a standing malady of the German body politic since 1198 and that the stability of the German monarchy depended largely upon the degree of cooperation achieved with the territorial princes, more especially with the prince-electors. On his return from his imperial coronation as Holy Roman emperor (1355) he promulgated, with the consent of the German assembly of estates or diet (1356), a basic constitutional document, known as the Golden Bull from its pendant gold seal (*bulla*). Charles's double objective was to minimize areas of dispute in future elections and to strengthen his ties with the electors. Unanimity among the electoral princes had always been difficult to attain; hence the validity of election by majority vote, a principle already set forth in the Declaration of Rhens, was reaffirmed. The territories of the lay electors were declared indivisible and heritable only by the eldest son. Thus, partitions of land by family agreement and consequent uncertainty concerning the holder of the electoral vote were eliminated. In conformity with ancient custom the archbishop of Mainz was to convene the electors and to request them to name their favoured candidate. He was to announce his own choice after the other electors had given their vote verbally, and so he could cast the deciding vote in the event of a tie. The election was to be held in Frankfurt, the royal coronation in Aachen.

Provisions of the Golden Bull

The membership of the electoral body was fixed at the traditional number of seven: the archbishops of Mainz, Cologne, and Trier, the count palatine of the Rhine, the king of Bohemia, the margrave of Brandenburg, and the duke of Saxony. When the throne was vacant the count palatine would be regent in south Germany and the duke of Saxony in the north; thus the long-standing papal claim to govern the empire during a vacancy was tacitly rejected. The question of papal confirmation of elections was ignored; neither Charles nor his electors were prepared to yield, but an open affirmation of their position would have been ill-received by the papacy, which had played a leading role in Charles's election.

The Golden Bull consolidated and extended the territorial power of the electors. Their right to construct castles, issue coinage, and impose tolls was confirmed. They could judge without appeal. Conspiracy or rebellion against them was deemed high treason. They were to meet the ruler once yearly as supreme advisory council on affairs of state. The formation of city leagues against them was specifically prohibited. On the basis of these enactments the Golden Bull has been called the Magna Carta of German particularism. The electors in their capacity of territorial lords were its chief beneficiaries; the rest of the princes were envious and strove thenceforth to acquire an equally large measure of territorial sovereignty.

Rudolf IV of Austria ordered his chancery to fabricate a series of imperial charters, including two from Julius Caesar and Nero, as evidence of his virtual independence of the empire. Charles IV submitted them for examination to the Italian humanist Petrarch, who declared the charters to be spurious. Rudolf took up arms and was bought off by the recognition of his claim to Tirol (1364).

The election of Charles's son Wenceslas (Wenzel) as king in 1376 (two years before Charles's death) was a striking example of the emperor's skill in securing the cooperation of the electors for his dynastic purposes. The election of an emperor's son as king of the Romans during the father's lifetime had not occurred since 1237; the prince-electors, in their anxiety to prevent any single dynasty from strengthening its grip on the succession, had checked all subsequent attempts. But unprecedented gifts, concessions, and a renewed prohibition of city leagues by Charles overcame the opposition of the electors. Pope Gregory XI had previously announced that the election would be invalid without papal confirmation. Charles, in concert with the electors, speeded the election and subsequent coronation of his son and then submitted an antedated request for confirmation to the pope, who countered these devious tactics by delaying confirmation; it was still under consideration at Gregory's death in 1378. The decline of the papacy during the Great Schism (1378–1417) precluded the vigorous assertion of its right of confirmation, which became a mere formality and was subsequently tacitly abandoned.

Election of Wenceslas

Decline of the German monarchy. Charles IV's power was based primarily upon the territorial possessions of the house of Luxembourg, which he greatly extended by the purchase of the electorate of Brandenburg (1373). The German monarchy was a source of dignity and influence, but in terms of land and revenue it was outranked by Charles's hereditary domains in the east and northeast. The Golden Bull, replete with privileges to the electors, attacked none of the fundamental problems of the monarchy: dwindling crown lands, slender revenues, lack of an army and of an expert bureaucracy.

The financial problem was acute and of long standing. The succession of disputed elections between 1198 and 1257 had compelled the various claimants to purchase support by grants of royal land and revenues; the attempt by Rudolf of Habsburg to recover possession of crown lands alienated since 1245 had been opposed by his electors, who were unwilling to set an example by surrendering their own considerable acquisitions. At every election the votes of the princes had been secured by the grant or pledge of royal rights and property; thus, every king began his reign with a financial millstone round his neck and could attain freedom of action only by the possession or acquisition of extensive dynastic territories. The system of pledging crown lands involved the transference of the land and its revenues to the creditor. These revenues did not reduce the original debt, and the alienation tended to become permanent. The imperial cities (*Reichsstädte*) had been heavily taxed by Rudolf, and before his acquisition of Austria they had furnished the bulk of his revenue. His less provident successors had pledged them in a few cases to the local territorial princes and had thus lost the right of taxation. Charles IV carefully cultivated his dynastic revenues from Bohemia, but he lavishly expended crown assets in Germany to expand his family possessions. His financial exploitation of the cities for purely dynastic purposes naturally stiffened their resistance to taxation. By 1400 the annual revenues from all the German crown possessions averaged only 30,000 florins.

Financial problems

The enforcement of the public peace, a taproot of royal power in other countries, had long since slipped from the hands of the German monarchs. The German monarchy possessed no executive officials comparable with the English sheriff or justice of the peace, and it was diverted from its guardianship of law and order by recurrent conflicts with the papacy and by its absorption in purely dynastic matters. Consequently, the proclamation and enforcement of the peace fell into the hands of regional associations of cities and of the individual territorial princes. Thus the monarchy was prevented from using its function as defender of the public peace as an entering wedge to invade the jurisdiction of the municipalities and the territorial lords.

In sum, the German rulers were being gradually deprived of their triple role of feudal suzerains, defenders of the church, and keepers of the peace. The sweeping privileges granted to the princes in 1220 and 1231 had undermined

their position as feudal suzerain. Their bitter struggles with the papacy cast doubt on their credibility as protectors of the church. They allowed their powers as guardians of the public peace to slip into the hands of others.

The continued ascendancy of the princes. By Charles IV's death in 1378, the division of Germany into loosely defined territorial principalities had reached an advanced stage.

Division into territorial principalities

Southern Germany. In south Germany the dissolution of the Hohenstaufen duchy of Swabia gave territorial predominance to the Habsburgs, whose original possessions lay in Alsace, Breisgau, the Vorarlberg, and Tirol. Rudolf's acquisition of Austria and Styria (1282) had more than doubled the Habsburg patrimony and established its centre of gravity in eastern Germany. The Habsburg's rivals and neighbours to the north, the counts of Württemberg, had combined with the Swabian nobles to foil the attempt of Rudolf to revive the defunct duchy of Swabia for one of his sons (the counts, insatiably acquisitive and the inveterate enemies of the cities of the region, were finally raised to ducal status in 1495). The margraves of Baden were chiefly preoccupied with the southward expansion of their territory on the upper Rhine at the expense of the independent small nobles and cities of Swabia.

These three large entities contained lesser lordships, which were in constant danger of absorption by marriage, purchase, or feud. Bavaria, granted to the house of Wittelsbach as a duchy in 1180, was strengthened by the acquisition of the Palatinate in 1214; but subsequent testamentary partition restricted this important gain to the Upper Palatinate.

Central Germany. In central Germany the margraves of Meissen of the Wettin dynasty thrust steadily eastward and received the electorate of Saxony in 1423, when the Ascanian line of electors died out; in the west they obtained Thuringia (1263) and clung to it tenaciously despite repeated royal attempts to oust them by claiming it as a vacant fief. The landgraves of Hesse, though surrounded by powerful neighbours, contrived to make modest territorial gains at the expense of the Wettin dynasty and the archbishops of Mainz. East and south of Hesse, the Rhine–Main region was a land of great ecclesiastical princes: the archbishops of Mainz, Trier, and Cologne; the bishops of Speyer, Worms, Würzburg, and Bamberg; and the wealthy abbots of Fulda and Lorsch. It abounded in counts of the second rank, dominated by a great secular prince, the count palatine of the Rhine. The area contained four electorates and was therefore of crucial political importance.

Northern Germany. In north Germany the dukes of Brunswick dissipated their strength by frequent divisions of their territory among heirs. Farther east the powerful duchy of Saxony was also split by partition between the Wittenberg and Lauenburg branches; the Wittenberg line was formally granted an electoral vote by the Golden Bull of 1356. The strength of the duchy lay in the military and commercial qualities of its predominantly free population. But the vigour of its eastward expansion into the Slav lands beyond the Elbe tended to diminish its involvement in the internal politics of the Reich.

Eastern Germany. In eastern Germany the duchy of Mecklenburg, Germanized by a steady stream of immigrants, was drawn deeply into Scandinavian affairs and in 1363 provided Sweden with a new royal dynasty in the person of Albert of Mecklenburg. The electorate of Brandenburg, purchased by Charles IV and bequeathed to his second son, Sigismund, was dominated by a disorderly and rapacious nobility. Sigismund granted this dubious asset in 1415 to his faithful ally Frederick, burgrave of Nürnberg. The kingdom of Bohemia remained the durable territorial core of the Luxembourg dominions, and its silver mines at Kuttenberg, under German supervision, vastly increased crown revenues. The Slav population resented increasingly the economic and cultural influence of the German minority, and this created antagonisms profoundly disturbing to the monarchy.

Continued dispersement of territory. Inside the various territories the consolidation of the princely authority was far from complete. The principalities were often ragged in outline and territorially dispersed because of the accidents of inheritance, grant, partition, and conquest. Everywhere lesser nobles disputed the power of the prince and formed associations in defense of their rights and fiefs. In the ecclesiastical princedoms the ascendancy of an archbishop or a bishop was contested by the cathedral chapter, which had become a preserve of the nobility. The self-governing cities fought to protect their chartered liberties and drew together in formidable leagues to resist princely encroachment. Thus the princes, trying to enforce their authority, tended to consolidate the opposition and to excite potential or open hostility.

Opposition to princely power

In this crucial struggle the great secular potentates impaired their own strength by persisting in the Germanic custom of dividing their territory among their sons instead of transmitting it intact to the eldest. By 1378 the Bavarian lands of the house of Wittelsbach were shared between three grandsons of Louis IV. In 1379 the wide possessions of the Habsburgs were partitioned by family agreement between Albert III and his younger brother Leopold.

The ecclesiastical princes, vowed to celibacy and elected by their cathedral chapters, could not hand on their lands to their descendants. Still, their policies and aspirations were not much different from those of the secular princes, and most of them managed to install their relatives in rich canonries and prebends.

1378 TO 1493

Internal strife among cities and princes. The electors had voted for Wenceslas reluctantly during Charles IV's reign, fearful that the monarchy might become a perquisite of the house of Luxembourg. Most of the other princes shared their concern over the continued ascendancy of the dynasty.

Wenceslas. Wenceslas (ruled 1378–1400) inherited a variety of problems, which grew after his father's statesmanlike hand had been removed. Wenceslas' habitual indolence and drunkenness, vices that increased as he grew older, excited the indignation of his critics. His prolonged periods of residence in Bohemia betrayed his lack of interest in German affairs and allowed the continuous friction between princes, cities, and nobility to develop into open warfare.

Conflict between princes and cities

The collision of princes and cities was prompted by vital issues of long standing. The flight of the rural population from servile tenures on the land to the free air of the cities reduced the labour force and impaired the revenues of territorial lords. Others who stayed on the land accepted the protection and jurisdiction of the neighbouring city as "external" citizens (*Ausbürger, Pfahlbürger*) and thus withdrew themselves and their land from seignorial control. Only the most powerful cities (*e.g.,* Nürnberg, Rothenburg) were able to extend their extramural territory to a substantial degree by force, but all strove to expand the area of their jurisdiction at the expense of local lords, partly to prevent village industries from competing with the city guilds.

A second major issue was the insistence of territorial lords on imposing tolls on city merchandise in transit through their possessions. In theory, tolls on road and river traffic were exacted in return for the protection of merchants and their goods, but the multiplication of toll stations hampered trade and provoked innumerable disputes, which often culminated in the seizure of merchants and merchandise by exigent lords.

The third and immediate cause of the crisis lay in the financial policy of Wenceslas himself. His Bohemian revenues, though large, were strained by the great sums payable to the electors in return for his elevation to the kingship. Hence he attempted to tap the resources of the imperial cities by demanding heavy taxes, and he threatened to mortgage recalcitrant cities to the neighbouring princes, their capital enemies.

On July 4, 1376, an alliance of 14 imperial cities of Swabia was formed under the leadership of Ulm and Constance for mutual protection against unjust taxes and alienation from the empire. The Swabian League counted 40 members by 1385 and was linked with similar coalitions in Alsace, the Rhineland, and Saxony. Wenceslas' initial hostility to the league faded as its membership in-

creased, and in 1387 he gave it his verbal and unofficial recognition. He feared to offend the territorial princes by extending full recognition; further, a clause of the Golden Bull had declared all city leagues to be illegal. Thus he temporized and awaited the outcome of the approaching trial of strength between cities and princes. On Aug. 28, 1388, the princes of Swabia and Franconia routed the largely mercenary forces of the Swabian League at Döffingen, near Stuttgart. The stipendiaries of the Rhenish League were put to flight by the count palatine Rupert II near Worms on November 6.

The cities triumphantly withstood the ensuing siege operations, but their economy was injured by the forays, ambuscades, and blockades instituted by the princes. The protracted campaigns also exhausted the financial resources of the princes. When Wenceslas intervened in 1389, both parties were ready for peace. At the Diet of Eger (May 2) he ordered them to desist and declared the city leagues to be dissolved. The contestants complied. The princes were satisfied with the prospective disbandment of the cities, and the cities feared the consequences of further resistance, but Wenceslas' opportunism was relished by neither side. The princes disliked his political flirtation with the cities, and the cities resented his final championship of the cause of the princes.

Wenceslas' alienation of cities and princes

Wenceslas' early gestures of support for the cities rankled with the electors, who in 1384 and 1387 discussed the advisability of replacing him by an imperial vicar or regent. Wenceslas, however, learned of the plan and conveyed his opposition; nor could the electors unite on their choice of a regent. Some electors turned to a more drastic solution—his deposition. In 1394 Rupert II and the archbishop Frederick of Cologne considered the election of Richard II of England but failed to win the support of their electoral colleagues. In the following year, however, Wenceslas' elevation of Gian Galeazzo Visconti, imperial vicar of Milan, to the status of duke was assailed as a dismemberment of the empire and enabled the electors to act as the indignant defenders of the integrity of the Reich against a wasteful and profligate king. Wenceslas attempted to conciliate the princes by appointing his younger brother Sigismund as German regent (1396). But the Milanese issue enabled Rupert and Frederick to enlist the support of the archbishops of Mainz and Trier for their proposed deposition of Wenceslas. The death of Rupert in 1398 occasioned some delay. But at length the electors compiled a lengthy series of charges against the king, and in September 1399 they openly proclaimed their intention of deposing him.

At this critical stage further proceedings were temporarily checked by serious differences concerning the choice of Wenceslas' successor. The favoured candidate of the Rhenish electors was the count palatine Rupert III, himself an elector. But another elector, Duke Rudolf of Saxony, and a powerful group of north German princes contended that the electors could not raise one of their own members to the kingship. The Golden Bull had declared otherwise, but Rudolf held his ground and declined to participate in the subsequent proceedings. On June 4, 1400, the four Rhenish electors invited Wenceslas to Oberlahnstein to consider measures for the reform of the empire and threatened to release themselves from their oath of allegiance if he failed to appear. The king's efforts to rally support for his cause were utterly fruitless, and he decided to stay in Bohemia. On August 20 Archbishop John of Mainz, on behalf of the four electors, publicly proclaimed the deposition of Wenceslas as an unfit and useless king and freed his German subjects from their allegiance to him. On the following day the three archbishops elected Rupert in Wenceslas' stead. Rupert's consent to his election was presumed to furnish the necessary majority required by the Golden Bull.

Deposition of Wenceslas

Rupert. Rupert (ruled 1400–10) lacked the skill and resources necessary to revive the drooping power of the German monarchy. His title was not beyond dispute while Wenceslas lived, and the territorial princes and cities were therefore slow to acknowledge him. Pope Boniface IX, maintaining that only a pope might legally depose a German monarch, withheld his approbation of Rupert. An expedition against Wenceslas (1401) failed before the walls of Prague. Rupert then embarked upon an Italian expedition (1401–02), hoping to obtain the imperial crown from the pope and thus dispel the cloud of uncertainty that hung over his title. The enterprise was crippled by lack of financial means, Boniface's conditions were exorbitant, and Rupert returned to Germany without the coveted imperial coronation. Fortunately, he had little to fear from Wenceslas, who was fully occupied in protecting his Bohemian throne from the machinations of his ambitious younger brother Sigismund. Far more dangerous was the degeneration of Rupert's relations with the Rhenish electors. In 1405 he offended Archbishop John of Mainz by refusing him military aid in his war against Hesse and Brunswick. Consequently the archbishop united all the enemies of Hesse and Brunswick in the League of Marbach, which included 18 imperial cities. Rupert contended that coalitions of cities were prohibited by the Golden Bull, and he denounced the league as illegal. The dispute was arrested by the mediation of the archbishop of Cologne, but the memory rankled. Rupert's prospects darkened still further in 1408, when he lent his support to Pope Gregory XII against the cardinals who wished to summon a general council to end the Great Schism in the church. The archbishops of Mainz and Cologne and the vast majority of the German prelates favoured the conciliar solution and strongly approved the policy of the cardinals. Wenceslas shrewdly followed suit and in return received assurances from the cardinals that the future general council would recognize him as German king. The powerful proconciliar party in the German church proceeded to agitate openly for the restoration of Wenceslas to the throne. The threat of civil war, however, was averted by Rupert's death on May 18, 1410.

Sigismund. On the death of Rupert the movement for the reinstatement of Wenceslas immediately lost headway. The Rhenish electors, having deposed Wenceslas 10 years previously on ground of his unfitness, could not reelect him without admitting their inconsistency. Nonetheless, the House of Luxembourg was powerful and would assuredly throw its full weight against any non-Luxembourg candidate to the German throne. The four electors agreed on the expediency of selecting Rupert's successor from the Luxembourg dynasty but disagreed on the choice of candidate. Rupert's successor, the count palatine, and the archbishop of Trier elected Wenceslas' brilliant but unreliable brother, Sigismund, at Frankfurt on Sept. 20, 1410. Eleven days later, the archbishops of Cologne and Mainz elected Wenceslas' turbulent and treacherous cousin, Jost of Moravia. Jost died the next year, and Wenceslas agreed to accept Sigismund on condition that he himself retained the title of German king. But Sigismund ignored the reservation and assumed the disputed title. Wenceslas' protests were greeted with indifference in Germany and quickly died away. A second election of Sigismund at Frankfurt (July 21, 1411) gave him an ample majority and removed all doubt concerning the validity of the previous election.

Sigismund was energetic, versatile, and intelligent; but long experience never blunted his rashness in rushing into new projects, and his financial incapacity never ceased to astonish his contemporaries. His pursuit of personal power and dynastic possessions was unceasing and was conducted with complete unscrupulousness. His kingdom of Hungary and his later acquisition, Bohemia, were his primary concerns, and the interests of Germany were constantly set aside in their favour. The disastrous reigns of his predecessors, Wenceslas and Rupert, had emphasized Germany's basic problems: the weakness of the monarchy, the friction between princes and cities, and the unchecked growth of lawlessness and disorder. During his long reign (1410–37) Sigismund appeared less and less frequently in Germany and did little to correct these evils.

Sigismund's character

The Hussite controversy. Sigismund's prolonged absences were caused in great part by the explosive Hussite controversy in Bohemia. The Czech church in Bohemia had long retained a marked individuality and much autonomy in its liturgy. This independent temper in ecclesiastical affairs was being slowly fused in the late 14th century with a rising sentiment of nationality among the Czechs. The upsurge of feeling took the negative form of a grow-

ing hatred of the German minority, which dominated the towns by virtue of its economic power and cultural influence. The luxury and immorality of the Bohemian clergy were castigated by a series of religious reformers such as Conrad of Waldhauser, Thomas of Štítný, John Milíč of Kroměříž (Kremsier), and Matthew of Janov. The teachings of Conrad and Milíč assumed a strongly puritanical tinge; in opposition to the wealthy sacramental church with its external means of grace they held up the ideal of the primitive church in a condition of apostolic poverty and the exclusive authority of the Bible as the foundation stone of faith and belief. These three movements met and intermingled in the person of Jan Hus.

Jan Hus. A graduate in divinity of Charles IV's foundation, the University of Prague, Hus was appointed incumbent of the Bethlehem Chapel in Prague (1402) and immediately attracted wide attention by his sermons, which were delivered in Czech in accordance with the foundation charter of the chapel. In 1403 he strongly defended a number of extracts drawn from the religious writings of the Englishman John Wycliffe. Czech opinion in the university solidly supported Hus, but the more numerous German masters carried the day, and the teaching of the controversial extracts was forbidden. Similarly, when Pope Gregory XII's cardinals rebelled against the pope and demanded a general council to terminate the schism in the papacy (1408), the Czech members of the university aligned themselves with the cardinals, while the Germans stood with the pope. On matters of general policy the masters of the university voted by "nations," of which there were four: Bohemian, Bavarian, Saxon, and Polish, the last consisting in fact largely of Germans. Thus the Germans controlled three votes, the Czechs only one. When King Wenceslas reversed the proportion by decree, the German masters and students seceded to found their own university at Leipzig, and the mutual enmity deepened.

In 1410 Hus was excommunicated by Archbishop Zbyněk of Prague but refused to appear at the papal court for judgment and continued to preach. Two years later he protested against the sale of indulgences and was placed under papal sentence of excommunication. The city of Prague was subjected to a papal interdict. Hus left the capital at Wenceslas' request but preached throughout the land and vastly enlarged his following. From the lower Czech clergy who popularized Hus's doctrines, the masses learned that the German minority were intruders, foes of Bohemia and of the true religion. Lesser nobles who had lost their lands by mortgage or purchase to the prosperous German burghers of the cities were readily converted. The more self-interested members of the upper nobility were attracted by Hus's proposed reduction of the Czech church to apostolic poverty, which would bring the rich territorial possessions of the higher clergy within their grasp.

The Council of Constance

The rising ferment in Bohemia disquieted the heir apparent, Sigismund, and he intervened with the suggestion that Hus should expound and justify his opinions to the Council of Constance (1414–18), recently convened to heal the schism in the papacy. Hus accepted, and Sigismund furnished him with a comprehensive safe conduct. The conciliar commission that examined Hus focused the debate on two issues: the unauthorized Bohemian practice of extending communion in both kinds (bread and wine) to the laity, and the points of agreement between Wycliffe and Hus. Hus declined to retract his Wycliffite opinions until they were refuted by Holy Writ, and thus he defied the authority of the council in matters of doctrine. He was arrested on Nov. 28, 1414, and died at the stake on July 6, 1415. Sigismund's protests against the breach of his safe conduct were silenced by the argument that excommunicates automatically lost imperial protection.

The Hussite wars. The death of Hus enshrined him at once as a martyr and a national hero in the memory of his followers among the Czechs. They raised a storm of denunciation against Sigismund and expressed their resentment by widespread attacks on orthodox priests and churches. The Catholics retaliated in kind, and Bohemia was in a state of civil war when the death of Wenceslas (Aug. 16, 1419) brought Sigismund to the tottering throne. The new king's talent for conciliation and compromise was useless in the heated religious atmosphere. Pope Martin V urged him on against the Hussites and promised him imperial coronation as his reward. Under his prompting, Sigismund raised a motley host in Germany and launched it into Bohemia under the banner of a papal crusade (March 1, 1420). But the invaders were thrown back from the walls of Prague, and on July 7, 1421, Sigismund was declared deposed by the Bohemian assembly of estates. The shock of defeat forced Sigismund to attempt a fuller mobilization of German resources. Under the traditional system, princes and cities had been allowed to fix at their own discretion the quota of men provided by each when a royal campaign was in prospect. Naturally, both estates used their discretionary power to reduce contributions to a minimum. In 1422, however, Sigismund himself fixed the strength of the contingents demanded from the individual princes and cities throughout Germany. The response was disappointing. In 1426 the king raised his requirements, but to no effect. Hence the yearly campaigns against the Hussites were waged largely by mercenary armies. To meet the rising costs, the Diet of Frankfurt was persuaded in 1427 to vote a general tax, the so-called Common Penny. But there was little enthusiasm in Germany for the crusade; massive evasions of payment occurred, and the strength of local feeling hampered the coercion of defaulters.

Hussite raids into Germany

In 1429–30 the irrepressible Hussites swept through Saxony, Thuringia, and Franconia in a destructive foray. Sigismund, exploiting the general alarm, reverted to the older system and demanded contingents from each prince and city. The response improved, and a large army invaded Bohemia, only to meet complete disaster at Taus (Domažlice in modern Czechoslovakia) in 1431. It was evident that the veteran Hussites could not be crushed by force. Sigismund therefore welcomed the opportunity to transfer the problem of reconciling the Hussites with the church to the Council of Basel (1431–49). The Hussite extremists, the Taborites, were inflexible. They condemned the hierarchical system of church government and affirmed the priesthood of all true believers. Hence the council conducted its long and arduous negotiations with the majority party among the Hussites, the Calixtines or Utraquists, who were prepared to accept the grant of communion in both kinds as a basis of settlement. The Utraquist nobles annihilated the protesting Taborites at the Battle of Lipany (May 30, 1434), made peace with the council by the Compact of Iglau (July 5, 1436), which conceded them communion in both kinds, and reunited with the Roman Catholic church. The Utraquist nobles extracted far better terms from Sigismund as the price of their recognition. He agreed to accept the guidance of Czech councillors in governmental affairs, to admit only Czechs to public office, to grant an amnesty for all offenses committed since the death of Wenceslas, and to allow the Czechs a large measure of autonomy in their civil and religious life. It is unlikely that the slippery Sigismund intended to honour these pledges, but they cleared the way for his triumphant return to Prague in August 1436.

Effects of the wars on Germany

In Germany, the Hussite threat had clearly revealed the inadequacies of the existing financial and military systems, but the incentive to press Sigismund's reforms to a successful conclusion faded when the Hussite peril was scotched by negotiation. The general apathy was demonstrated in 1434, when Sigismund proposed to the princes a land peace embracing the whole of Germany. The abolition of private wars and feuds by such a peace was undeniably a paramount necessity. The princes themselves, however, were among the chief offenders against law and order, and their nominal approval of the plan deceived no one. Sigismund himself, increasingly absorbed in crucial negotiations with the Hussites, did not persevere, and the project gathered dust in the imperial archives. The impulse he gave to the cause of reform did not, however, fade entirely, though Sigismund did not live to see the sequel. His death on Dec. 9, 1437, terminated the tenure of the German throne by the House of Luxembourg and opened the door of opportunity to the Habsburg dynasty.

The Habsburgs and the imperial office. *Albert II.* In the

absence of a male heir, Sigismund had named his son-in-law Albert of Habsburg, duke of Austria, as his successor. Albert was able and vigorous, and the union of the territories of the two dynasties enabled him to exert considerable leverage in German politics. Albert declared his neutrality in the current dispute between Pope Eugenius IV and the Council of Basel on the subject of conciliar sovereignty and thereby evaded an issue on which the electors were strongly divided; thus, on March 18, 1438, he was unanimously elected at Frankfurt. The electors attempted to elicit from the new king an understanding that he would grant privileges to his subjects only with their advice and consent. They also submitted a project for the division of Germany into four new administrative units (*Kreise*) in which the enforcement of the land peace would be entrusted to captains of princely rank. Albert judged that the princes were seeking to enlarge their power and influence under the guise of introducing reforms for the common good. The German cities also doubted the impartiality of the princes as custodians of law and order. Both proposals were therefore stillborn. The king hastened from Frankfurt to defend his kingdom of Hungary, endangered by Turkish raids on Siebenbürgen (Transylvania in modern Romania). The campaign was brought to a premature close by the death of the king on Oct. 27, 1439.

Frederick III. Albert II had left only an infant son, and the leadership of the House of Habsburg passed to his cousin Frederick, duke of Styria. Inside the electoral college the duke was vigorously supported by his brother-in-law Frederick of Saxony and was elected unanimously on Feb. 2, 1440. The choice of Frederick tightened the hold of the Habsburgs on the German kingship. It also brought to the throne a ruler who, absorbed in dynastic concerns and in astrology, had no more than a passing interest in Germany. Under the absentee government of Frederick III, the feuds among the princes and the collisions between the princes and the cities developed into savage wars accompanied by widespread ravaging and pillage. All paid lip service to the need for peace; but who was to enforce it? Was it to be enforced by the monarchy, which lacked power and executive machinery? Was it to be enforced in the courts of the princes, whose judicial impartiality was suspect? Were complaints against the princes to be heard and decided in the king's court (*Hofgericht*)? Or must they be adjudicated by the council (*Hofrat*) of the prince concerned? The right to enforce peace effectively was a major source of power to the holder; hence the struggle between Frederick and the princes was long, bitter, and inconclusive.

The Turkish threat

These issues were brought to a head by the rapid westward progress of the Ottoman Turks after their victories at Varna on the Black Sea (1444) and at Kossovo in Serbia (1448). The Habsburg kingdom of Hungary and Frederick III's own duchy of Styria lay full in the path of the invaders. In 1453 the fall of Constantinople extinguished the Eastern Empire and aroused fears in Germany that the Western Empire would meet the same fate. The king used the opportunity to demand financial aid against the Turks from the diet, the German assembly of estates. Under the leadership of the princes, the diet reminded him that Germany's capacity for defense was weakened by the current internal anarchy. In 1455 six electors proposed to the king the establishment of an imperial court of justice in which all three estates (electors, princes, and cities) should be represented. Frederick dismissed the scheme as an attempted invasion of his authority and stubbornly maintained his disapproval in a series of stormy interviews.

In time an increasing number of princes became convinced that reform would make no significant progress until Frederick was removed. As early as 1460 the Wittelsbach princes urged his deposition in favour of George of Podèbrady, the able and resourceful king of Bohemia. To check the danger, Frederick began to dole out reforms with a sparing hand. In 1464 he consented to make the court of the treasury (*Kammergericht*) independent of his person, to staff it with representatives of the three estates, and to extend its jurisdiction into fields other than financial. It was the acquisition of Austria in 1463 on the death of his brother Albert that finally proved his undoing. The unruly Austrian nobility early took the measure of Frederick and thereafter disregarded his authority. On the east and south the duchy was imminently threatened by the expanding kingdom of Hungary under its land-hungry ruler Matthias Corvinus. The southern borders of the Habsburg lands were also ravaged by the Turks. Frederick's continuing irresolution and passivity encouraged Matthias Corvinus, who had already seized a portion of Bohemia, to launch a campaign against Austria. The Austrian nobility made no move against him, and Vienna fell to him in 1485. Frederick fled to Germany and made pitiful appeals for help to the princes. His misfortune provided the party of reform with a long-awaited opportunity. Led by Berthold of Henneberg, the able and resolute archbishop of Mainz, they pressed the aging and afflicted Frederick to relinquish the kingship in favour of his son Maximilian. Solaced somewhat by the assurance of a Habsburg succession, he gave a reluctant acquiescence, and Maximilian was elected on Feb. 16, 1486. Frederick retained the title of emperor, held since his imperial coronation at Rome in 1452. But he played no part in the government of Germany, and his death on Aug. 19, 1493, passed almost unnoticed.

Developments in the individual states to about 1500. *The princes and the Landstände.* In the various principalities the outcome of the struggle between the territorial princes and the assembly of estates (*Landstände*) was not fully decided by 1500. The vigour of the conflict arose partly out of the contrasting conceptions of government held by the protagonists. The secular princes looked upon their lands as private possessions that could be divided by agreement among their sons and drew little distinction between their private and their public revenues. The three estates regarded themselves as the corporate representatives of the whole territorial community and maintained that actions by a prince affecting their interests and privileges should be subject to their consent. They therefore opposed the partition of the territory by family pact among the princes' sons. The inadvisability of breaking up the principalities into petty territorial lordships was at length conceded by the more prudent princes. By 1500 the rulers of Bavaria, Brandenburg, Saxony, and Württemberg had accepted the principles of territorial indivisibility and primogeniture.

Dispute over taxation

In financial matters the imposition of extraordinary taxes (*Notbeden*) remained the crucial issue between the princes and the estates. The mounting cost of war and administration outstripped the ordinary revenues of the ruler, plunged him deeply into debt, and compelled him to seek financial aid from the estates with increasing frequency. In the absence of a clear distinction between public and private revenue, the estates often contended that the deficit was a private debt of the prince and disclaimed responsibility. Needy princes were thus forced to buy temporary solvency by concessions that later shackled them in their dealings with the estates. The estates regarded the *Notbede* strictly as an occasional emergency tax and insisted that it should be reasonable in amount. Indeed, the estates of Bavaria and Brunswick extracted from their respective princes in the course of the 14th century a formal recognition of their right of armed resistance to extortionate taxation. Similarly, any prince who broke his agreements with the estates was subject to the right of resistance. Thus the aims of these territorial assemblies were mixed. They sought to preserve the privileges of the three orders, to restrain the power of the prince, and to limit taxation. They were, however, also actively interested in good government, and the more enlightened rulers usually issued their ordinances only after consultation with the estates.

The princes proceeded against these powerful and often turbulent bodies with great caution. They persistently demanded that territorial assemblies convene only at the summons of the prince. They discountenanced the widespread conviction that absentees from the assembly were not bound to pay the taxes that it voted. In consequence, the peasants, who were not represented except in the Swiss cantons, Baden, Friesland, and Tirol, remained in the grasp of the princes' tax collectors. In these directions the princes had generally made notable advances by 1500, but in the vital matter of the *Notbede* they were still obliged to bargain with the estates as equals. They had

nowhere attained their ultimate objective: to transform the tax into a regular imposition voted automatically by the estates on demand.

Beyond the confines of the assembly of estates the attempts of the princes to curb their overmighty subjects aroused vigorous resistance. The noble vassals, proud and unruly, readily combined against any prince who sought to tamper with their liberties. Wise rulers deflected the nobles' energies into useful channels by employing them as stipendiaries. Hence even the most powerful princes—the Habsburgs in Austria and the Hohenzollerns in Brandenburg—proceeded circumspectly, and the difficult task of bringing the nobility to heel was far from completed in 1500. The cities of the princely territories defended their independence no less stubbornly. The princes revoked their charters, influenced municipal elections, and forbade the cities to associate in self-defense. The struggle was most intense in the north and east, where the Hohenzollern dynasty of Brandenburg emerged as the chief foe of municipal freedom. In 1442 the elector Frederick II ("Iron Tooth") crushed a federation of Brandenburg cities and deprived its leader, Berlin, of its most valued privileges. In the Franconian possessions of the dynasty, Albert Achilles of Hohenzollern waged a destructive war (1449–50) against a city league headed by Nürnberg. He suffered a resounding defeat in a pitched battle near Pillenreuth (1450). The elector John Cicero took up the battle 38 years later, when the cities of the Altmark in west Brandenburg refused to pay an excise tax on beer voted by the assembly of estates. He discomfited the cities in the ensuing "Beer War" and radically revised their constitutions to his own advantage. On the other hand, the great cities of south Germany, enriched by the Italian trade, were more than a match for the local princes: the Wittelsbach dukes of Bavaria were decisively worsted by Regensburg in 1488.

The "Beer War"

The growth of central governments. Between 1300 and 1500 the organs of central government in the territorial states became more specialized and diversified. The parent body was the advisory council (*Hofrat*) of high nobles and ecclesiastics, whom the prince consulted at his discretion. Its business was not differentiated, and there was no division of labour among the councillors. It met at the summons of the prince and did not convene at regular intervals. Its membership was not fixed, and some advisers did not attend except at special invitation. Others were regional councillors who attended the prince only when he appeared in their locality. A body so unspecialized and fluctuating was ill-adapted to cope with the increasingly complex problems of central government. Hence in the 14th and 15th centuries a professional element of "daily" or permanent councillors was introduced. They were usually legists, trained in Italy or in the newly founded universities of Prague (1364), Vienna (1365), Heidelberg (1386), Rostock (1419), and Tübingen (1477). They were well versed in Roman law, which, with its centralizing and authoritative precepts, provided a congenial climate for the growth of the powers of the territorial princes everywhere save in Saxony and Schleswig-Holstein, where the ancient customary codes were deeply rooted. Financial administration, which required specialized skills, was placed under the direction of a separate department of government, the chamber (*Hofkammer*). An inner ring of favoured advisers, the privy council (*Geheimrat*), was also instituted to counsel the prince on affairs of state. The besetting weakness of the new administrative structure was financial. Few princes followed the example of the Hohenzollern dynasty in drawing up an annual budget and requiring financial officials to submit regular accounts to the government. On the positive side, chanceries gradually created a common German language, which Luther later used to spread his message.

German society, economy, and culture in the 14th and 15th centuries. *Transformation of rural life.* Despite the impressive advance of trade and industry in the later Middle Ages, German society was still sustained chiefly by agriculture. Of an estimated population of 12 million in 1500, only 1.5 million resided in cities and towns. Agriculture exhibited strong regional differences in organization. The more recently settled areas of the north and northeast were characterized by great farms and extensive estates that produced a surplus of grain for export through the Baltic ports. The south and southwest was a region of denser population, thickly sown with small villages and the "dwarf" estates of the lesser nobility. In the northeast the great landlords, headed by the Knights of the Teutonic Order, tightened their control of the originally free tenants, denied them freedom of movement, and ultimately bound them to the soil as serfs. In the south the heavy urban demand for grain chiefly benefited the larger peasant proprietors, who sold their surplus production in the nearest town and used their gains to acquire more land. The lesser peasantry, with their smaller holdings, practiced chiefly subsistence farming, produced no surplus, and therefore failed to benefit from the buoyant urban demand. The frequent division of the patrimony among heirs often reduced it to uneconomically small fragments and encouraged an exodus to the cities. On the other hand, landless day labourers who survived the Black Death in the mid-14th century were able to command higher wages for their services.

Serfdom in the north

In south Germany the strain of transition in rural society was heightened by the policies of the landlords, lay and ecclesiastical. Confronted by labour shortages and rising costs, many landlords attempted to recoup themselves at the expense of their tenants. By means of ordinances passed in the manorial courts they denied to the peasantry their traditional right of access to commons, woods, and streams. Further, they revived their demands for the performance of obsolete labour services and enforced the collection of the extraordinary taxes on behalf of the prince. The peasants protested and appealed to custom, but their sole legal recourse was to the manorial court, where their objections were silenced or ignored. Ecclesiastical landlords were especially efficient, and peasant discontent assumed a strong anticlerical tinge and gave rise to the localized disturbances in Gotha (1391), Bregenz (1407), Rottweil (1420), and Worms (1421). Disturbances recurred with increasing frequency in the course of the 15th century on the upper Rhine, in Alsace, and in the Black Forest. In 1458 a cattle tax imposed by the archbishop of Salzburg kindled a peasant insurrection, which spread to Styria, Carinthia, and Carniola. In Alsace the malcontents adopted as the symbol of revolt the *Bundschuh,* the wooden shoe usually worn by the peasants. They also formulated a series of specific demands, which included the abolition of the hated manorial courts and the reduction of feudal dues and public taxes to a trifling annual amount. On these fundamental points there was little room for compromise, and the outbreaks were stifled by the heavy hand of established authority. But the rigours of repression added fuel to peasant discontent, which finally burst forth in the great uprising of 1524–25 (see below).

The nobility. The lesser nobility included two distinct elements. The imperial knights (*Reichsritter*) held their estates as tenants in chief of the crown. The provincial nobility (*Landesadel*) had lost direct contact with the crown and were being compelled by degrees to acknowledge the suzerainty of the local prince. The imperial knights had been extensively employed by the Hohenstaufen emperors in military and administrative capacities and were chiefly concentrated in the Hohenstaufen possessions in Swabia, Franconia, Alsace, and the Rhineland. With the extinction of the Hohenstaufen dynasty they lost their function and rewards as a nobility of service. The revenues from their small estates sank in purchasing power as prices rose. Caste prejudice prevented them from seeking an alternative role in trade or industry. Resentful of the decline in their fortunes and fiercely independent, they clung grimly to their remaining privileges: exemption from imperial taxes and the right to indulge in private war. They stubbornly resisted the persistent attempts of the princes to reduce them to subject status, and in Trier and Württemberg especially they were given valuable aid by the provincial nobles. For purposes of defense or aggression the imperial and provincial knights combined freely in powerful regional leagues, usually directed against the local princes or cities. In the course of their chronic feuds with the cities, many knights

became mere highwaymen. Many others, who had been forced to sell their estates or who were encumbered with debts, took service in Germany or Italy as mercenaries (*Soldritter*). In east Germany the knights, though equally unruly, were far more affluent. The knightly estate (*Rittergut*) was larger and produced a profitable surplus for export. The knights sat in the assembly of estates, and taxation by the prince required their consent. They were therefore well entrenched against the encroachments of princely power.

Urban life. Urban society in 15th-century Germany was concentrated in some 3,000 cities and towns. About 2,800 of the total were extremely small, with populations varying from 100 to 1,000. Of the remainder, no more than 15 cities contained more than 10,000 inhabitants. In this restricted group three were preeminent. Cologne reached its peak in the 13th century with a population of 60,000, but sank to 40,000 by 1500 following internal disputes, expulsions, and steady emigration. In 1500 Augsburg was the most populous German city, with a resident population of 50,000. Third place was held by fast-growing Nürnberg, which counted 30,000 souls. The social unity of the citizen body had been most marked in the 13th century, when the guilds joined the dominant patrician families (*Geschlechter*) in wresting the right to form an independent city council (*Stadtrat*) from the lord of the city. In the 14th century the guild masters, methodically excluded from the council by the patrician oligarchy, broke into open revolt in Speyer (1327), Strassburg (1332), Nürnberg (1348), and elsewhere. In its economic aspect the ensuing conflict embodied an attempt by the guildsmen as industrial producers to free urban industry from the tight control exercised by the merchant patriciate. By 1500 the guilds almost everywhere had gained varying degrees of representation in the city council.

The guild system

In the meantime the guilds themselves had become increasingly oligarchical and exclusive as the established masters restricted the entry of new members in order to reduce competition. The ascent of journeymen and apprentices to the rank of master was obstructed by the imposition of excessive fees, and in many guilds membership became virtually hereditary. In consequence, the journeymen began to associate in fraternities of their own to press their demands for higher wages and a shorter working day. The masters denounced the fraternities as illegal, compiled blacklists of leading agitators, and formed intercity associations to enforce low wage rates. The day labourers and casual workers outside the guild structure had no protective organization and suffered heavily in periods of economic depression. The surviving tax records of the German cities, though not wholly reliable guides, nevertheless suggest wide extremes of wealth and poverty. In late 15th-century Augsburg, 2,985 of a total of 4,485 households (66 percent) were recorded on the tax rolls as exempt from taxation on the ground of insufficient means. At the other extreme stood the enterprising and prosperous business dynasties of the Fugger and the Welser. Not all wealthy citizens lacked public spirit, however, and hospitals, almshouses, and charitable foundations multiplied within the city walls. The spreading problem of mendicancy was combated by stringent legislation against ablebodied beggars in Esslingen (1384), Brunswick (1400), Vienna (1442), Cologne (1446), and Nürnberg (1447).

The decline of the church. The vigour and assertiveness of secular society in Germany was exercised increasingly at the expense of the clergy and the church. Among the upper clergy more than 100 archbishops, bishops, and abbots were temporal rulers. The prelates were usually sons of the nobility and did not allow election to church office to interfere with their aristocratic ardour for war and territorial acquisition. They were expert in the accumulation of benefices and were notoriously lax in the performance of their spiritual duties. Their influence was freely used to advance their kinsmen and partisans among the greater and lesser nobles, who dominated the cathedral chapters and ruled the abbeys. The monasteries were filled with monks and nuns who were distinguishable from the lay aristocracy only by a nominal celibacy. Among the secular princes, the rulers of Austria, Brandenburg, and Saxony wrested a right of appointment to a fixed number of bishoprics and lesser church offices from the papacy, which had been gravely weakened by the schism and the conciliar movement of the 15th century. All lay and ecclesiastical princes imposed heavy extraordinary taxes on the clergy. The steady invasion of the church by secular interests was also exemplified by the moral and material condition of the lower clergy. The Black Death of 1348–49 had decimated the ranks of the more dedicated priests, who ministered to their plague-stricken flocks instead of seeking safety by flight. The new recruits who rushed into holy orders were often self-seeking and spiritually unqualified. As the inflow continued, the problem of clerical unemployment and inadequate stipends attained greater proportions. Many were compelled by need to accept illpaid livings. Others obtained no benefice at all and lived precariously as chantry priests or as itinerant chaplains. Their moral and intellectual defects were bitterly assailed by church reformers and by an increasingly well-informed laity. Many pious Christians, especially in the cities, began to turn away from the priesthood in their search for spiritual comfort and to seek relief in mysticism or in lay associations practicing a simple, undogmatic form of Christianity.

Trade and industry. The most impressive achievements of the German economy between 1200 and 1500 lay in trade and industry. German trade benefited from the Hundred Years' War between France and England, which diverted northbound Mediterranean merchandise from the customary Rhône valley route to the eastern Alpine passes; from the fierce internal warfare between the Italian city-states, which weakened their supremacy in long-distance trade; and from the rapid economic development of "colonial" eastern Europe between the Baltic and the Danube. The north German trade was chiefly based on staple commodities such as grain, fish, salt, and metals; but the south German merchants, in their capacity as middlemen between Italy and the rest of Europe, had taken the lead by 1500. They combined trade and industry in the great Ravensburg Trading Company (1380–1530), which produced and exported Swabian linen and laid the foundation of the fortunes of the Höchstetter, Herwart, Adler, Tucher, and Imhof families. The most important independent concern was that of the Fugger, whose founder, Hans Fugger, began his career as a linen weaver in Augsburg. The Fuggers' accumulated profits provided capital for moneylending and banking, which they conducted with the aid of business techniques borrowed from the more advanced Italians. The wealth and prosperity of Germany, which Machiavelli remarked on in 1512, stood in sharp contrast to its political and military weakness, a disparity that contributed significantly to a profound sense of malaise and discontent on the eve of the Reformation.

The Ravensburg Trading Company

Cultural life. In the absence of a strong centralized monarchy to act as a focus, German culture continued to be regional in character and widely diffused. The mysticism of Meister Eckhart, Johann Tauler, and Heinrich Suso, which commanded all men to look for the kingdom of God within themselves, flourished chiefly in the cities of the Rhineland, where lack of diligent pastoral care forced Christians to call upon their own inner resources. In the same region social and moral satire attained an urgent and vivid realism. Sebastian Brant (1458–1521), born at Strassburg, spared no class in his epic on human stupidity, the *Narrenschiff*, or *Ship of Fools*. But it was in the thriving cities of south Germany, as yet little affected by Italian humanism, that late Gothic culture reached magnificent heights in art, architecture, and sculpture. Albrecht Dürer, born in Nürnberg in 1471, challenged his generation with his evocative engraving of "Melancolia I" in which a brooding figure with closed wings sits idly amid a chaos of scientific instruments and meditates on the futility of human endeavour. In architecture the hierarchical elaboration of the late Gothic style maintained its ascendancy and even made a notable conquest in Italy with the construction of the great cathedral of Milan, begun in 1387. The sculptured carvings of Tilman Riemenschneider (*c.* 1460–1531) in the castle of Würzburg revealed the anxiety, the deep piety, and the religious sensibility of

Christian men engaged upon a spiritual pilgrimage that was to continue to the Reformation and beyond. His work was the pinnacle of a great flowering of sculpture, one of the greatest in German history. (C.C.B./L.G.D.)

Germany from 1493 to *c.* 1760

REFORM AND REFORMATION: 1493–1555

The empire in 1493. The reign of Maximilian I (1493–1519) was dominated by the interplay of three issues of decisive importance to the future of the Holy Roman Empire: the rise of the Austrian House of Habsburg to international prominence, the urgent need to reform the empire's governing institutions, and the beginnings of the religious and social movement known as the Reformation. The accession of the dynamic and imaginative Maximilian to the German throne aroused in many Germans, and in particular among humanists, expectations of a time when the old imperial idea—the vision of the empire as the political expression of a united Christendom in which the emperor, as God's deputy, rules over a universal realm of peace and order—might become a reality. Since the extinction of the Hohenstaufen dynasty in 1254, imperial authority had been in disarray; as weak emperors had become absorbed in struggles against foreign and domestic, secular and ecclesiastical rivals, real power in the empire had moved toward the governments of territorial states and independent cities. From their first appearance on the historical scene (briefly from 1273 to 1308, then from 1438 in a nearly unbroken line until the dissolution of the empire in 1806), Habsburg rulers had fostered imperial unity. But they had been notably unsuccessful in creating agencies for its attainment, partly because they were assiduously working to build up a power base for their own house. This, in turn, brought them into conflict with European antagonists, chiefly France. The long reign of Maximilian's father, Frederick III (1440–93), was regarded by nationalists as lamentable in its inattention to the problems pressing on the empire. The solutions proposed for them were subsumed under the name of "reform," a highly charged word that acquired enormous additional force in the 15th century when the conciliar movement and its lay and clerical proponents exerted pressure for religious renewal. Maximilian's arrival on the throne thus generated a surge of anticipation, expressed in an outpouring of agendas for restructuring what was then coming to be called the "Holy Roman Empire of the German Nation."

Imperial reform. Inevitably, perhaps, Maximilian's performance with regard to the empire disappointed. But he was successful in significantly extending the dynastic might of his family. He took over the duchy of Tirol with its vast mining resources. Ambitious marriage alliances spread Habsburg entitlements west and east: in 1496 Maximilian's son Philip wed Joan, the daughter of the king and queen of Spain, thus linking Habsburg Austria to Spain and the Netherlands (the future Charles V was born of this union in 1500); and in 1516 Maximilian's grandson Ferdinand was betrothed to the heiress of Hungary and Bohemia. But these connections only escalated Maximilian's internal and external problems. In foreign politics his ventures ended, for the most part, in calamities for the empire. Switzerland was lost (1499), several Italian campaigns were repulsed, and an attempt against the duchy of Burgundy brought the hostility of France and led to the fall of Milan. Even the imperial crown eluded Maximilian: his advance on Rome for this purpose was halted, and he had to be content with the self-bestowed title of "Roman Emperor Elect."

These reverses strengthened a reform party among leading members of the empire's estates (Stände), especially their spokesman, the archbishop-elector of Mainz, Berthold von Henneberg. Given the long rivalry between emperor and estates, it goes without saying that their respective plans for reforming the empire diverged on crucial points of direction and control. The estates, acting from their—to them entirely legitimate—sense of the prerogatives of particularism, favoured a central administration responsive to them; the emperor, to the contrary, insisted on organs subservient to him, modeled on bureaucratic agencies recently established in Burgundy and Austria and shored up by the authority-enhancing principles of Roman law.

At the Imperial Diet held in the city of Worms in 1495 the estates, whose members had begun to see themselves as the authentic representatives of the whole country, prevailed. The four reform measures adopted on this occasion were in large part intended to limit the emperor's powers. An "Eternal Peace" outlawed private feuds, and steps were taken, agreed to by the emperor, to implement this pacification. An Imperial Chamber Court functioned as supreme tribunal for the empire, most of its judges being named by the estates. An empirewide tax, the "Common Penny," was imposed, the collection of which fell to the estates. To these measures was added, in 1500, an Imperial Governing Council to monitor, under the domination of the electors, the empire's foreign policy. Maximilian was able to impede the operation of this body (it was reestablished in 1521), but overall the emperor's reform objectives had come to nothing.

The historical judgment on this failure has swung between scorn for Maximilian's imperial "fantasies"—for whose realization a reorganized Germany was to furnish the means—and respect for his grasp of the country's perilous geopolitical situation, and between sympathy for the estates' single-minded pursuit of the imperatives of regional state building and disdain for the parochialism of their political vision. Most historians have found Maximilian's actions distorted by a species of romantic idealism (he in all seriousness proposed to Bayzid II that Turks and Christians should settle their differences by the rules of the tournament; he also cherished for a time the hope of becoming pope as well as emperor). What is clear is that the reform effort so briskly launched at the beginning of his reign resulted in permanent confrontation between sovereign and estates, a posture in which neither party could outmaneuver the other and every matter of policy required tenacious negotiation. In light of this permanent tug of war, the notion of a universal realm ruled by a "German" emperor does seem fantastic, though it never lost its ideological force. No one in the empire possessed plenary authority, but real power was shifting to territorial rulers and their bureaucratic agents and to the magistrates of larger cities. This shift was to be of the greatest importance in determining the direction in which the Reformation established itself in Germany.

The Reformation. The Reformation presents the historian with an acute instance of the general problem of scholarly interpretation—namely, whether events are shaped primarily by individuals or by the net of historical circumstances enmeshing them. The phenomenon that became the Protestant Reformation is unthinkable without the sense of mission and compelling personality of Martin Luther. But in social and intellectual conditions less conducive to drastic change, Luther's voice would have gone unheard and his actions been forgotten. Among the preconditions—which are the deeper causes of the Reformation—the following stand out: (1) Everyone agreed that the Roman Catholic church was in need of correction. The lack of spirituality in high places, the blatant fiscalism, of which the unrestrained hawking of indulgences—the actual trigger of the Reformation—was a galling example, and the embroilment in political affairs were symptoms of corruption long overdue for purgation. While the church continued to be accepted as the only legitimate mediator of divine grace, denunciations of its abuses, perceived or actual, became more strident in the decades before 1517. (2) A subtle change, moreover, had been occurring in people's religious needs and expectations, leading to demands for a more personal experience of the divine. Failing to meet this aspiration, the church was widely, if diffusely, rebuked for its unresponsiveness. (3) More focused criticism came from the Christian humanists, an influential group of scholars bent on restoring the fundamental texts of Western Christianity. Led by Desiderius Erasmus, the most renowned biblical scholar of the time, these men held the Catholic church up to the spiritual ideals for which it claimed to stand and, finding it wanting, set the principle of Evangelicalism against the church's secular-

Causes of the Reformation

ized ambitions. (4) Still more fatefully, by 1500 the church had come under attack from European rulers whose administrative, legal, and financial hegemony could not be completed in their respective states without domination of the ecclesiastical sector. In the empire, as elsewhere, the trend in ecclesiastical politics was toward state churches (*Landeskirchen*), in which governments, using "reform" as a pretext, gradually gained, while church authorities lost, a large measure of control over clerical properties, personnel, and functions. The Reformation was the culmination of this process, which, in the empire, took place in nearly all princely territories and in most independent cities where governments brought the administration of the church under political direction. (5) In Germany this development was facilitated by an ancient feudal custom entitling a landlord to extend "protection" to churches located on his estates. Over this "owner's church" (*Eigenkirche*) he enjoyed the right of patronage, allowing him to appoint incumbents and manage properties. In the course of extending their sovereignty, territorial princes took over this right to patronage and fashioned of it the legal basis on which, in the Reformation, they assumed full control over the administration of the church. (6) In every segment of German society, but particularly among the poor, voices were being raised against injustice and exploitation. Wide disparities in income and discriminatory laws in cities as well as the deteriorating standard of living of small peasants and agricultural labourers caused riots and uprisings, which by the early 1500s had become endemic.

These, then, were the forces driving events toward a crisis. In the first decade of the 16th century they coalesced into a powerful surge of religious, social, and political agitation, for which "reform" (of church and society) was the code word. Ironically, Luther, who was to channel this agitation into the Reformation, had, until his emergence as a national figure in the 1520s, nothing to do with it. For him one issue alone mattered: the imperative of faith. His personal path to the Reformation was an inner search for religious truth, to which his conscience was his guide.

Martin Luther

When he wrote his Ninety-five Theses against indulgences in October 1517, Luther was an Augustinian friar, a preacher in the Saxon city of Wittenberg, and a theology professor at the university founded there in 1502 by the elector of Saxony, Frederick III, called "the Wise." His ambitious father had pushed him toward a career in law, but in 1505 the fervently devout Martin entered a monastic house. His order, that of the Augustinian eremites, was a strict reform congregation dedicated to prayer, study, and the ascetic life. Deeply troubled by the question of justification—of how a human being, a sinner, may be justified (saved) in God's sight—Luther found no comfort in monastic routine and turned to an exploration of the sources of the Christian faith, notably St. Paul and St. Augustine. His intellectual promise having been recognized, he was sent by his order to study theology at Erfurt and Wittenberg. He gained his doctorate in 1512 and commenced his teaching of the Bible in Wittenberg that same year. According to his own account, it was during his close reading of Paul's Epistle to the Romans, while preparing to give a course of lectures on that text, that he discovered what struck him as the solution to the problem posed by the huge gap between human sin and divine grace. Justification is not earned as a reward for human effort through good works (a position Luther now attributed to a misguided and misguiding Roman church). To the contrary, human beings are justified without any merit of their own by God's freely given and prevenient (*i.e.*, coming before any worthy human deeds) grace, through faith, which is a gift of God. This is the meaning Luther found in the crucial passage in Romans 1:17: "For in it [*i.e.*, the gospel] the righteousness of God is revealed through faith for faith: as it is written, 'He who through faith is righteous shall live.' " "Righteousness"—*justitia* in Latin—does not refer, Luther now saw, to God's activity as judge but to the justifying righteous condition he effects in the human sinner, a condition expressing itself as faith. The momentous consequences of this theological insight, which Luther appears to have taken as a unique discovery but which had in fact been espoused by a score of theologians before him, were not then apparent to him. They asserted themselves powerfully, however, once he began to lecture and preach on the—for him—paramount themes of salvation by faith alone (*sola fide*) and exclusive reliance on Scripture (*sola scriptura*). It was the indulgence controversy of October 1517 that brought it all into the open.

Few other issues could so clearly have exposed the gulf that separated this ardent friar from an urbane and pragmatic church. The indulgence offered in Saxony in 1517 had its origin in two purely financial arrangements. First, Popes Julius II and Leo X needed funds for rebuilding Saint Peter's Basilica in Rome; second, Bishop Albert of Hohenzollern, forced to buy papal dispensations in order to gain the archbishoprics of Mainz and Halberstadt, agreed to promote indulgences in his domains, half the income from which was to go to Rome, the other half to him and his bankers. For Luther, the issue turned not so much on the outrageous venality of this deal as on the indulgence itself. Truly contrite sinners do not desire relief through an indulgence (which is a remission of temporal punishment to be performed following absolution); they crave penance. This is the gist of Luther's argument in the Ninety-five Theses, which he sent to his ecclesiastical superiors to persuade them to abandon the indulgence sale. (The story that he nailed a copy of the theses to the door of the castle church in Wittenberg may be the invention of a later time.)

Luther intended no defiance with this action. He intervened as a priest on behalf of his flock and as a conscientious theologian against a corrupting church. But the public reaction to the theses (he had written them in Latin, but they were soon translated and printed) made it evident that he had touched a nerve. Encouraged by expressions of support and goaded by opponents, Luther became more outspoken, harsher in his criticism of the church, and more focused in his attacks on the papacy. By 1520 he was well on his way to becoming the spokesman for Germany's grievances against Rome. A pamphlet he published that year, "Address to the Christian Nobility of the German Nation," urged the empire's secular rulers to reform a church that would not set its own house in order. Popes and prelates are not sacrosanct, he argued; they may be brought to justice. As every Christian can read the Bible for himself, papal claims to interpretive authority are a vain boast. Luther prodded the German princes to consider the state of the church and to reform it for the sake of the faith. In this way Luther drew out, albeit reluctantly, the full consequences of his principle of "salvation by faith alone." No church was needed to act as God's agent; grace was available without mediation. No priest, not even the pope, has special powers, for, so Luther argued, all human beings are priests, made so by their faith. It is scarcely surprising that a bull of excommunication against him (*Exsurge domine*) issued from Rome in June 1520.

Imperial election of 1519 and the Diet of Worms. At any other time the Lutheran matter would probably have ended there. But 1521 was no ordinary moment in the empire's history. When he died in 1519, Maximilian had not succeeded in having his grandson and heir, Charles, designated his successor. King of Castile and Aragon since 1516, and suzerain also over Habsburg apanages in the Netherlands, Naples, and central and eastern Europe—not to mention the Spanish possessions in the New World—Charles posed a formidable problem for the electors. Should they choose a man whose vast resources might well empower him to centralize authority in the empire when this would limit their hard-won autonomy? Should they elect a prince whose international commitments would entangle Germany in European conflicts? The only other viable candidate was the king of France, Francis I, who recognized in the rapidly growing Habsburg ascendancy a serious threat to his own power. But, though Francis made tempting promises, the electors, in the end, opted for Charles—not, however, without drawing from him a 36-point "electoral capitulation" in which he swore to uphold the estates' prerogatives. While in so choosing the electors strove to secure their own privileges, the German public tended to see in the new monarch a fulfillment of national

aspirations—mistakenly, as it turned out, but giving rise, at the time, to patriotic expressions of hopes for a brighter future. Coinciding with these events in the political realm and gaining resonance from them, the Lutheran matter was inevitably drawn into the political vortex, thereby acquiring the features of a national cause.

The Diet of Worms of 1521

Delayed by disturbances in his Spanish domains, Charles reached Germany late in 1520. He was crowned king in Aachen, assuming at the same time the title of Roman emperor elect. He then proceeded to Worms, where he was to meet with the German estates in early 1521. By then no other issue counted as much on the agenda as the Lutheran affair. Acting out of what appears to be a blend of conviction and political expediency, the estates' leaders, prompted by Frederick the Wise, demanded that the diet reopen Luther's case by allowing the excommunicated friar to speak before the estates. Unable to resist, the emperor issued a safe-conduct, and Luther traveled to Worms, treated on his way to so extravagant a public acclaim that he must have felt like a national hero. On April 17 and 18 he stood before the emperor and representatives of the estates. He refused to revoke his views on justification and other points of theology, reiterated his denial of ultimate authority to pope and church councils, and—when pressed—asserted the principle of individual responsibility in matters of faith: "As long as my conscience is held captive by the words of God, I cannot and will not revoke anything, for it is dangerous, and a great peril to salvation, to act against conscience. God help me. Amen." These words, and all else that transpired, were broadcast to the country in scores of pamphlets in which Luther was cast in the role of the man sent by God to cleanse the German church. While an imperial edict condemned his teachings and placed him and his adherents under the ban, Luther himself was offered a refuge in Frederick the Wise's castle in the Thuringian forests, the Wartburg. There, fretting over his enforced absence from events, he turned to the translation of the New Testament, while the movement, of which he was now the acknowledged head, gained momentum.

Analysis of what happened at Worms reveals the first signs of what was to become a fateful split in the self-perception of this movement. The estates saw it as a means of promoting ecclesiastical reform; on questions of faith, they were willing to compromise. Luther, on the contrary, tolerated no distinction between points of faith and aspects of church practice, and on his understanding of the former he stood rock solid. He and his political patrons were thus pursuing different ends; for the present, however, his support in official circles was, for whatever reason, substantial.

Luther's hold on the general public was even more impressive than his hold on the political leadership. Historians are not unanimous in explaining how this friar from the German east, utterly obscure only five years earlier, could have gained such a following by 1521 that few governments would enforce the ban against him, knowing that they would face strong resistance if they tried to do so. It is not clear whether the general population, still largely illiterate, understood the doctrine of solifidianism well enough to make it the object of informed choice. More likely, the German populace took Luther not as the preeminently religious prophet he was but saw in him their best hope of achieving an amelioration of the many troubles vexing them in their respective stations, not only in religion but also, and perhaps mainly, in their social condition. Against these conditions, the products of deeply rooted legal and institutional structures, Luther—or so people seem to have understood him—raised the standard of evangelical morality: all things were to be judged by Scripture and God's law.

The issue of social reform

Scholars have often pointed out that this view is less attributable to the Wittenberg reformer than to another theologian and preacher then beginning to be active in the Swiss city of Zürich, Huldrych Zwingli. Luther explicitly rejected the use of the—to him purely spiritual—New Testament as a norm for social reform. But Zwingli affirmed it. On the other hand, Luther's vehemence and forceful rhetoric were so compelling that, rightly or wrongly, his name came to be fused with the general hope of improvement in human affairs. His partisans, proficient in their use of the media of print and woodcut illustration, helped shape this conviction by furnishing propaganda for a strong popular drive toward Lutheranism. This drive from below, further nourished by traditional anticlerical sentiments, was met from above by the eagerness of territorial and urban governments to utilize Lutheran ideas as legitimation for the extension of political control over the church. Thus, by the mid-1520s, a number of German cities and states had formally turned Lutheran, meaning that they had severed their legal and administrative ties to Rome and its prelates and were building new ecclesiastical institutions and framing new doctrines.

The revolution of 1525. The events of the revolutionary years 1524–25 underscored the urgent need to establish Lutheran church organizations. In its own time, and often since then, these events were labeled a "peasant rebellion"; but modern scholarship has made it clear that the insurrection was far more than a series of uprisings by rural bands. The tens of thousands of peasants drawn into the movement, some of them massed in major military actions, were a symptom of the general unrest that had gripped Germany since the middle of the 15th century. This unrest was ultimately caused by demographic pressures with attendant economic and political dislocations. The particular demands pressed in the 1520s—mitigation of fiscal and labour burdens imposed on peasants by their lords, autonomy for village communes, and relief from high taxes—had been voiced before. New was the linkage of these demands with the grievances of restive urban groups also protesting exploitation and disenfranchisement and their formulation as an agenda of social reform on the principle of Christian communitarianism. This ideological redirection of old patterns of resistance could not have occurred without the impetus of the Reformation, specifically the incendiary preaching in towns and villages of evangelical pastors who represented Lutheran and Zwinglian ideas as solutions to the problems at hand. The clearest evidence of the Reformation's impact on the shaping of what some modern scholars call "the revolution of the common man" are the Twelve Articles drawn up for the Swabian peasantry by an evangelical cleric and associate of Zwingli's. Article three of this document asserts that "The Bible proves that we are free, and we want to be free," while another article claims for every congregation the right to choose its minister; these articles are a strong indication of how vital the principle of Biblicism had become to people preeminently concerned with worldly life.

In the regions involved—Franconia, Swabia, the Upper Rhine and Alsace, Thuringia, and Tirol—large forces of peasants attacked castles, monasteries, and some cities. News of these actions encouraged discontented urban groups to rise against their oligarchic town governments, and for a while it looked as though a united revolutionary front of ordinary—*i.e.*, nonprivileged—people might be forged. Manifestos and lists of articles abounded; there was talk everywhere of judging things by "God's law" (meaning the gospel), and some groups even laid plans for a "Christian association" across regional and town–country lines. But before long the forces of authority won the upper hand, and the insurrectionaries were put down with the ferocity customary in those days. The war's final stage was dominated by Thomas Müntzer, a visionary theologian with a message of social deliverance for and by the poor. The defeat of his forces at Frankenhausen in May 1525 marks the final victory of the old order over the would-be new dispensation.

Thomas Müntzer

Luther was heavily implicated in this turnabout. Realizing that his words and deeds had served to encourage popular action against legitimate rulers, he sought to separate himself drastically from the movement, going so far as to urge the rulers' warriors to "cut [the peasants] down, hit them, choke them wherever you can." His brush with revolution confirmed Luther in the rigorous segregation he made between "two realms," the worldly and the spiritual, the respective laws and ideals of which must not be confounded, as, he claimed, the revolutionaries had done. The gospel, he said, cannot be used as a standard

for governing in the world, which has its own rules and ways of justice, many of them, he acknowledged, unfair and blameworthy. Luther's separation of worldly and evangelical values, soon made binding law by Lutheran governments, brought an abrupt end to the early phase of the Reformation, during which events seemed to many to be moving toward a sweeping transformation of social, as well as religious, structures. At the same time, Lutheran governments read 1525 as a lesson on the need to control subjects, concluding that preaching in particular, and religious behaviour generally, must be controlled. To accomplish this purpose, new laws and bureaucracies were set up everywhere.

Lutheran church organization and confessionalization. The 1525 revolution was but one of several upheavals worrying the authorities. Three years earlier a group of imperial knights led by Franz von Sickingen had declared feud on the archbishop of Trier, claiming to derive from Scripture their right to despoil Roman Catholic prelates. The ensuing "Knights' War" was quickly crushed. But about the same time a disturbance broke out in Wittenberg where, during Luther's exile in the Wartburg, a group of reforming Spiritualist activists forced the city council to abolish many traditional Catholic practices. Upset by this rash move, Luther intervened to reverse it. But this incident, and the knights' attack, caused consternation among the heads of government, who feared loss of control. Their anxiety was deepened by the spread from Switzerland in the mid-1520s of Anabaptism, a radical religious movement whose most distinctive tenet was adult baptism. The events of 1525 thus strengthened a long-ripening resolution that firm structures and clear doctrines were needed to reassert authority in a situation of drift.

The ability of Lutheran states to act on this resolution was facilitated by the impact of foreign affairs on the empire's internal politics. Far from seeing to the execution of the 1521 edict against Luther, Charles V left his brother Ferdinand in charge of imperial affairs and departed from Germany after the Worms diet to deal with the many problems besetting his far-flung interests. The most perilous of these was the war with France, which implicated the emperor in a constantly shifting balance of alliances with other powers and in a see-saw of military actions in which now he, now Francis I, was in the ascendant. Charles's victory at Pavia in 1525 led in turn to the formation of a coalition against him (the so-called "Holy League of Cognac"), intended to forestall Habsburg hegemony in Europe (a scenario to be replayed many times in the following two centuries). In 1526, therefore, Charles was in no position to dictate to the German estates on the Lutheran matter. Within a year, however, the situation turned in his favour. Spanish troops captured and plundered Rome in 1527, and by 1529 Charles was dominant once more, though it had become clear that neither warring party could bring the other to its knees. At the same time a potentially fatal danger loomed in the east where the Turks, under Süleyman I the Magnificent, began to aim their path of conquest at the Balkans and Hungary. The death of King Louis II at the Battle of Mohács in 1526 put Ferdinand of Habsburg (Charles V's brother) in line for the Bohemian and Hungarian crowns, thereby exposing the already overextended Habsburgs on a new front. By 1529 the Turks were moving toward Buda (captured in September of that year) and Vienna. Facing these perils, Charles concluded peace with France, sealing his triumph in the west with his coronation as emperor at Bologna. He then returned to Germany.

The Turkish advance

The events just described formed the larger political frame in which Lutheran church organization took place. Forced to solicit military aid from the estates in 1526, Ferdinand postponed implementation of the Worms edict, accepting a declaration by the Diet of Speyer of that year to the effect that every estate "will, with its subjects, act, live, and govern in matters touching the Worms edict in a way each can justify before God and his Imperial Majesty." This declaration gave Lutheran rulers the signal to proceed with their intended legal, administrative, financial, and liturgical reforms, and the years following 1526 saw the construction in every Lutheran territory of what amounted to a state church, headed by the ruling prince. In 1529 this process was interrupted when, following the emperor's military successes, Ferdinand demanded at a diet, also held in Speyer, that, pending a general council to decide the religious issue, Lutherans should end their separation. (It was the "protest" of a number of princes and cities against this abrogation of the earlier Speyer decree that attached to Lutherans the name "Protestants.") By then, Protestants were no longer a united party. Luther and Zwingli had met at Marburg in 1529 in an attempt to iron out differences, but they could not agree on the question of Christ's real presence in the Eucharist. While a few Lutheran princes prepared for military action, a compromise-minded group led by the humanist Philipp Melanchthon, who dreaded the prospect of fragmentation within Protestantism, drew up a moderate outline of Lutheran positions. These were presented for discussion at the Diet of Augsburg in 1530, which was attended by the emperor. The Augsburg Confession, which became a fundamental statement of Lutheran belief, assumed that reconciliation with the Catholics was still possible. This view was shared by Charles, who was pushing the pope toward the summoning of a general council to mend the religious split. Negotiations among theologians and politicians, however, came to nothing, and the end result was that, with the Augsburg Confession rejected, Lutheranism was outlawed again. The militant Protestant faction, led by Philip, landgrave of Hesse, now established a formal organization of resistance, the Schmalkaldic League (1531), and the empire moved toward armed conflict as Lutherans passed beyond the formation of a political party and became a military force as well.

Around this time much more rigid standards of religious orthodoxy and conformity were imposed. This development has been called "confessionalization," a concept used by some historians to define developments in the empire during the second half of the 16th century. Confessionalization completed the process, under way since the late Middle Ages, of meshing religious and church politics with the objectives of the state. Central to this process was the institution of a territorial religion that was based on an authorized declaration of doctrines (a "Confession") binding on all subjects and implemented by an established church responsible to the ruler (or, in city states, to the magistrates). Tending toward exclusiveness and therefore intolerance, this system contributed to the warlike turn taken by events after 1530. More important in the long run, confessionalism promoted a social drive, also long under way, toward the inculcation of discipline and order in public and private as well as in religious and civic affairs. Through catechisms, schooling, family and welfare legislation, norms for work, and standards for personal life, state and church attempted to restructure society in accordance with the goals of what has since been called the Protestant temperament. Success was slow in coming and never more than partial. But there is no doubt that, through the confessional process, Protestantism left a deep imprint on the German character.

Confessionalization

Religious war and the Peace of Augsburg. After the diet of 1530 Charles left Germany for more than a decade, occupied with troubles in the Mediterranean, the Netherlands, and, once again, France. In 1535 he campaigned against Tunis to subdue the Barbary pirates who, as the naval arm of the Ottomans and as corsairs and privateers, had been making navigation unsafe. Renewed war with France was temporarily halted in 1538 by a treaty meant to last 10 years, but in 1542 France struck again, allied now with major European powers, including the duke of Guelders and Cleves, whose lands were claimed by Charles as part of his Burgundian inheritance. The emperor's conquest of this duchy in 1543, which considerably broadened his power base, and the peace he concluded with France in 1544 (Peace of Crépy), followed by an armistice in 1545 with the Ottoman Empire, left him free at last to deal decisively with the German Protestants.

The emperor's policy toward religious deviants was guided by his concept of empire. The universal realm over which he hoped to reign faced external and internal threats; its mission of unity and order was assaulted by

infidels from without and by national rivalries and heresy from within. He had dealt with the first and second threats; now he turned his attention to the third. Protestantism had spread rapidly in Germany. More than a religion, it was, by the 1540s, a full-fledged political movement with a growing military capacity. The number of Protestant territories had recently grown to include, among others, Brandenburg, the Palatinate, Albertine Saxony, and the bishoprics of Cologne, Münster, Osnabrück, Naumburg, and Merseburg. In Philip of Hesse the Lutherans had an able political strategist. At least provisionally, pending the settlement of all religious issues by a general council, the Protestants had won grudging recognition of their right to exist. Such a council was actually summoned by Pope Paul III—though only upon repeated prodding by the emperor—but there were few signs that the Protestant states would submit. In 1545, therefore, Charles decided on war. He found a pretext in the capture, by Lutheran princes, of the duke of Braunschweig-Wolfenbüttel, a Catholic who had tried to reconquer the lands from which he had been expelled by his Lutheran subjects. Claiming that this capture violated imperial law, Charles opened the conflict in 1546, in which he was joined by Maurice, duke of Saxony, an ambitious Lutheran prince to whom Charles had secretly promised the Saxon electorship. The ensuing war fell into two phases, the first of which saw the emperor victorious (Battle of Mühlberg, 1547). Capitalizing on this strong position, Charles in 1548 forced the estates to accept an Interim, a temporary religious settlement on the emperor's terms. It was the political concessions Charles demanded from the estates, however—concessions that would have permanently limited their autonomy—that led to a resumption of war. Among the Protestants the lead was now taken by Maurice of Saxony, who had abandoned the emperor and had obtained material support from the new French king, Henry II, for fighting on the Protestant side. The resulting "Princes' War" was brief (1552–53) and inconclusive, and in 1555 a peace was signed at an imperial diet held, again, in Augsburg.

The Peace of Augsburg

The Peace of Augsburg closed one epoch of German history and opened another. It decided the religious issue but did so in a way bound to occasion future problems. It reinforced the princes' authority over their territories but failed to settle their relations with the emperor. Most important, it legalized Lutheranism, laying down the rule, later epitomized in the phrase *cuius regio, eius religio* ("he who governs the territory decides its religion"), that each estate—*i.e.*, each prince or city—could opt for either the Roman Catholic or the Lutheran religion (*jus reformandi*) and that this choice was binding on everyone under that ruler's jurisdiction. Only one faith could legitimately exist in a given state, and that faith had to be the ruler's and could be only Catholicism or Lutheranism; Calvinism, Zwinglianism, and Anabaptism were excluded. A subject unwilling to live by this choice was free to emigrate and take his belongings with him (a provision considered liberal at the time). Confiscated church properties could be kept by the governments that had taken them. An Ecclesiastical Reservation prevented ruling prelates from converting their lands along with them. These terms make it clear that the real winners of the war, and of the entire Reformation period, were the territorial princes, whose authority and power, which now encompassed the church, were greatly increased. It scarcely needs to be said that Luther, who had died in 1546, would not have approved of this outcome. As for the emperor, he abdicated in frustration and retired to a monastery in Spain, leaving his Spanish and Burgundian crowns to his son Philip and the empire to his brother Ferdinand. These two men, as Philip II and Ferdinand I, strong-minded Catholics both, were to play prominent roles in the period of Counter-Reformation and confessionalism that dominated Europe after 1555.

THE CONFESSIONAL AGE: 1555–1648

German society in the later 1500s. The changes caused by state building and the Reformation bestowed little real benefit on the lives of ordinary people. Historians agree that the later 16th century was, for many, a time of economic hardship and social stress. Rapid increase in population (the European population rose by more than half between 1500 and 1700) and, secondarily, the influx of precious metals from the New World were the main causes of an inflationary trend that spanned the entire century and reached painful stages in Germany in the 1590s and early 1600s. Grain prices were especially affected, with the result that an ever smaller share of the ordinary person's budget was available for the purchase of other products. This had several effects, which, at least in outline, are well documented. The quality of nutrition for all but the wealthiest became much worse than it had been in the late Middle Ages, when meat consumption was at an all-time high. Illness and epidemic disease were frequent as the nutritional deficiency was aggravated by a series of bad harvests, perhaps caused by unusually severe winters in the decades after 1560.

Cities and towns suffered loss of income as the market for their manufactured wares declined. In consequence, municipal guilds lost ground, not only economically but also politically, owing to the curtailment of their participation in urban policy-making. There were exceptions to this trend. Craftsmen specializing in the manufacture of luxury cloths and arms found lucrative markets at princely courts; but overall the position of artisans declined. Journeymen could no longer anticipate becoming masters. Artisans employed in traditional handiwork felt the pressure of the putting-out system favoured by early capitalism, whereby much production was moved from the town to the countryside. Division of labour was introduced, gradually transforming self-employed craftsmen into dependent workers.

In the agricultural sector, high grain prices and rising land values improved the lot of peasant proprietors, but the real beneficiaries were landowning nobles and urban patricians with investments in agriculture. Society was polarized by these developments. A minority of rich peasants lived amid struggling smallholders hard-pressed by landlords who maximized their profits by increasing labour and tax burdens (the period has been spoken of as that of a "second serfdom"), and in the cities an upper crust of merchants and landed aristocrats faced a proletariat, whole sections of which were pauperized by the end of the century. The populous cities, once the glory of Germany, began to play a smaller role, as economic troubles and the centralizing policies of territorial princes decreased their prosperity and sapped their political strength.

Social polarization

The highly visible contrast between rich and poor and the animosity of the weak against the wielders of influence created tensions among groups and classes. Political and economic power was more concentrated than ever before. Its new centres in Germany were the splendid courts of secular and ecclesiastical princes, whence it was distributed to favoured groups: the nobility, rising in importance again but finding its function limited to the service of rulers, and the upper bourgeoisie, shifting its loyalty from guild hall to palace. For ordinary people, administrative centralization and politically sanctioned Reformation had the effect of making their lives more rigid. A host of mandates flowed from centres of government, seeking to promote an ethic of order, productivity, and morality by shaping working and domestic activities as well as private habits and attitudes. These inroads caused resentment, and there is evidence of widespread resistance, most of it passive. Under these circumstances, the evangelical Reformation seems to have made but slight impact on the populace at large, whose effective religion continued to be a mixture of traditional Christianity and folk magic.

Most people were worse off near the end of the 16th century than at its beginning. The lot of women, in particular, had deteriorated. Around 1500 many German women had been at work in numerous urban occupations. But a century later they had been crowded out of all but the most demeaning trades as economic pressures, reinforcing ancient prejudices, eliminated them wherever they offered competition to male craftsmen. In this light, it is not surprising that the period from the 1580s to the 1620s also witnessed a surge of persecutions for witchcraft in Germany (mainly in the southwest and Bavaria). As

elsewhere, the witch craze in the empire seems to have been a reaction to the strains of a time of troubles, the actual causes of which, fairly clear now to historians, were hidden from contemporaries.

Religion and politics: 1555–1618. Four forces contended for supremacy in the Holy Roman Empire in the aftermath of the Peace of Augsburg. Lutherans—that is to say, Lutheran estates and governments—sought to extend the rights they had won in 1555 to parts of Germany still Roman Catholic. Calvinists, having been excluded from the Augsburg settlement, strove for recognition and made major territorial gains in the 1560s and '70s. Adherents of the old faith, invigorated by the Catholic Reformation issuing from Spain and Rome, attempted to turn back the Protestant advance by making common cause with strong governments. Habsburg emperors tried to serve the Catholic cause by weakening Protestant princes wherever possible and by holding the line against Protestantism in their dynastic lands. Political conflicts were constant under these circumstances and wars frequent, the empire's institutions being powerless to neutralize or channel these competing endeavours. Maximilian II from 1564, Rudolf II from 1576, and Matthias from 1612, though ardent Catholics, were preoccupied with the intertwined problems of retaining the loyalty of their dynastic lands and securing the eastern borders against the Turks. They had, in any case, been stripped of the ability to maintain order in the empire, as the Augsburg terms had placed public security under the supervision of the empire's administrative districts, which were controlled by the estates. The period leading up to the Thirty Years' War was therefore one of more or less constant strife in nearly all parts of the empire.

The second half of the 16th century introduced two new agents of change to this scene. The Catholic Reformation, operating mainly through the Council of Trent (1545–63) and the Society of Jesus, brought about major changes in Roman Catholicism. Trent produced authoritative definitions of dogma for the first time in the long history of the church, declared tradition to be, with the Bible, a source of revelation, reaffirmed the sacraments as mediators of grace, declared the church to be a hierarchical institution headed by the pope (against Luther's formulation of the "priesthood of all believers"), and issued a large number of reform mandates to meet, at last, the age-old charges of laxness and corruption. After the 1560s the Catholic Reformation's chief energies went to the implementation of the Trent decrees. Most effective in this endeavour were the Jesuits, a militant order founded by Ignatius of Loyola in 1534, pledged to strict obedience to the pope and to acting as the church's instrument for regaining ground lost to Protestantism. Germany was a major area of Jesuit activity; the order settled in Cologne in 1544 and later in Vienna, Ingolstadt, and Prague. In close collaboration with Catholic rulers, often as their confessors, the Jesuits embodied the activist phase of Catholic reform that goes under the name of Counter-Reformation.

The Counter-Reformation

On the Protestant side, this activism was represented by the Calvinists, who made so forceful an impact on German society in these decades that some historians have called their appearance a "Second Reformation." The Palatine Electorate went Calvinist when its ruler converted; later the "Reformed" creed (as its partisans named it, denying to other Protestant denominations the claim to have truly reformed the faith) established itself, among other places, in the electorates of Brandenburg and (for a time) Saxony, the territories of Hessen-Kassel, Nassau, Durlach, and Anhalt, and the cities of Bremen, Emden, and Münster. Unlike Genevan Calvinism, the Reformed religion in Germany coexisted easily with the autocratic territorial church; German princes, for their part, saw Calvinism as a far more aggressive theological and political weapon with which to wage the struggle for Protestant supremacy in the empire. Calvinist theology, with its emphasis on action in the world and its association of success with sanctification, even with election, was well suited, in the use made of it by German state churches, to act as an aggressive creed of social discipline; it also inspired the formation of militant parties pressing for recognition of Calvinism as a legitimate religion under the *cuius regio, eius religio* rule.

With the religious situation thus more inflamed than ever and the confessional and political issues inextricably intertwined, any incident might have triggered renewed conflict, which—given the competition for power in Europe among the Habsburg dynasties, France, England, and the Netherlands—was likely to lead to a general war. A series of incidents moved events toward the brink. In 1582 the archbishop-elector of Cologne, having converted to Calvinism, challenged the Ecclesiastical Reservation of the 1555 Augsburg treaty by holding on to his title, thus threatening to throw the majority vote in the College of Electors to the Protestants. In the "Cologne War" of 1583 he was expelled by Spanish troops, and Duke Ernst of Bavaria was chosen as his successor. Throughout the 1590s the incorporation of church properties by Protestant governments was a cause of litigation before the empire's courts, as Roman Catholic authorities sought to compel the return of everything confiscated since 1555; Protestant estates, in turn, made support for the emperor's war against the Turks dependent on further concessions.

The Habsburgs, meanwhile, were hampered in their advancement of the Roman Catholic cause by the growing mental incapacity of Rudolf II; indeed, much of the direction of affairs was transferred to his brother Matthias, who eventually succeeded him in 1612. A more serious undermining of Habsburg imperial power occurred in the dynastic lands. Rigorous Catholic reform occasioned peasant uprisings in Austria and resistance by nobles in Hungary and Bohemia (where Calvinism had made inroads among the ruling classes). In Hungary a nationalist party under István Bocskay forged an alliance with Lutheran princes and obtained support from the Turks. In Bohemia, in 1609, the estates extracted from the emperor a guarantee of religious freedom, the so-called Letter of Majesty.

At about the same time the city of Donauwörth was occupied by Bavarian troops, Duke Maximilian of Bavaria having been empowered by the emperor to "protect" the Roman Catholic minority there. Seeing this "Donauwörth incident" as a straw in the wind, Lutheran and Calvinist rulers formed a Protestant Union (1608), the answer to which was the Catholic League (1609), headed by Maximilian, the most resolute Catholic prince in the empire. Each party organized an army and allied itself with foreign powers, the Protestants with France and Bohemia, the Catholics with Spain. In this way the German struggle was both militarized and internationalized.

General war nearly broke out in 1609–10 over the Jülich-Cleves succession crisis. When the Roman Catholic ruler of these counties, which formed the strategically most important block of territories on the lower Rhine, died without issue, two Protestant claimants occupied his lands, aided not only by the German Protestant Union but also by France and England; they were, however, militantly opposed by Spain and the emperor. The assassination of Henry IV of France, who had been about to launch an invasion in support of the Protestant claimants, defused the crisis in 1610.

The Jülich-Cleves succession crisis

Peace was preserved, although not for long. The Bohemian situation finally precipitated the war. Because neither Rudolf II nor Matthias left legitimate heirs, the governance of the Habsburg dynastic lands fell to Archduke Ferdinand of Styria (later Emperor Ferdinand II), a ruthless counter-reformer who reduced the religious liberties granted to Bohemians under the Letter of Majesty. In response, the Bohemian estates in May 1618 mounted a protest in Hradčany (Prague Castle), which prompted a militant faction of deputies to throw two imperial councillors from a castle window ("defenestration" being a traditional Bohemian gesture of defiance). Ferdinand now prepared military action, while the Bohemian estates elected a Calvinist, Frederick V of the Palatinate, to be their king. As the alliances fell into place on each side, the stage was set for the sequence of large-scale military actions that constituted the Thirty Years' War.

Defenestration of Prague

The Thirty Years' War and the Peace of Westphalia. The Bohemian problem was resolved swiftly. Two Roman Catholic armies, the emperor's and the League's, converged on the kingdom, routing Frederick at the White

Mountain in November 1620 and replacing the regime of the estates in Bohemia with a system of "confessional absolutism" based on rigid Catholic conformity and political authoritarianism. At the same time, the Palatinate was conquered by Spanish and Bavarian troops, and the electoral title, forfeited by Frederick, transferred to Maximilian of Bavaria in 1623. In the Palatinate, too, the Counter-Reformation sought to bring Protestantism to an end. As the war spread into Hesse and Westphalia, and as Spain resumed its attack on the Netherlands, Catholic forces seemed near triumph. This prospect, however, renewed the old fear, as in the time of Charles V, of Habsburg hegemony; an anti-Habsburg alliance was therefore forged by France (where Cardinal Richelieu took charge of affairs in 1624), England (James I was father-in-law to the deposed Frederick V), the Netherlands, and Denmark (whose Protestant king, Christian IV, had extensive territorial interests in north Germany, now threatened by Catholic armies). In 1625 Christian commenced hostilities. He was opposed by a much-enlarged imperial force under the war's most flamboyant figure, Albrecht von Wallenstein, a military entrepreneur with a gift for mobilizing troops and supplying them through ruthless plundering. Wallenstein's plan was to wreck Dutch and English commerce in the Baltic by subduing all of north Germany and by dislodging the Swedish king, Gustav II Adolf, from Prussia (taken in the course of Sweden's war against Poland). By 1628 much of this objective was realized, Christian IV having been decisively defeated in 1626 in the Battle of Lutter am Barenberge and forced to make peace in 1629. To crown these successes, Ferdinand II issued in March of that year an Edict of Restitution, by which Protestant rulers were to restore to the church more than 500 bishoprics, monasteries, abbeys, and other ecclesiastical properties secularized since 1552.

But this impressive strengthening of the sovereign's power in the empire brought his foreign and domestic enemies together once more, the latter including now not only Protestants but also Roman Catholic estates concerned about their liberties. Subsidized by the Dutch and by France and allied with Saxony, Sweden entered the conflict in 1630, winning commanding victories at Breitenfeld (1631) and Lützen (1632) but suffering defeat at Nördlingen in 1634. This phase of the war was marked by unprecedented brutality; for example, in 1631, imperial troops massacred two-thirds of the population of Magdeburg, a city of 20,000 that had withstood a long siege.

The Magdeburg massacre

A way out of the long conflict appeared in 1635 when Saxony, Brandenburg, and other Protestant estates seeking to end foreign intervention joined the emperor in the Peace of Prague, which included the revocation of the Edict of Restitution. But in this, the war's final phase, the initiative passed to France, which, seeking to forestall Spanish preponderance on the Continent, offered large subsidies to Sweden and to German princes to enable them to fight on. Combined Swedish-French campaigns commenced in 1638, and a decade later foreign armies operated as far south as Bavaria, while the French held Lorraine and Alsace, which was important to France to prevent construction of a Spanish land bridge from the Netherlands to Italy.

By then most belligerents were exhausted. Several German princes had quit the war. Since 1644, representatives of the powers had been talking about terms, although military operations continued in hopes of improving bargaining positions. In 1648, finally, treaties were signed in Münster and Osnabrück (both in Westphalia) by agents of the emperor, the German estates, Sweden, and France as well as between Spain and the Netherlands. Fighting continued for some years—France and Spain did not conclude peace until 1659—but the war was at last winding down.

The Treaties of Westphalia

The Peace of Westphalia brought territorial gains to Sweden and France, awarded an electoral seat to Bavaria, and secured for Protestant rulers the church properties they had confiscated, based on the status quo of 1624. More important, it brought Calvinists into the religious settlement and established the independence of the Netherlands from Spain and of Switzerland from the empire. Most significant of all, it guaranteed the nearly unlimited territorial sovereignty of German princes, bringing to an end the last effort (until the 19th century) to turn the empire into a centrally ruled "modern" state. In this way 1648 sealed the fragmentation of the Holy Roman Empire into hundreds of autonomous political entities, most of them small. At the same time, it brought to an end the last major conflict in continental Europe in which religion was a salient issue; indeed, the war itself had demonstrated that reason of state was a stronger determinant of policy than faith. In declaring the religious situation fixed as of 1624, the treaty eliminated the *jus reformandi* as a cause for confessional change; if a prince converted, his land no longer converted with him. Religious pluralism and—albeit grudgingly—coexistence were now the norm.

The war's social and economic cost is difficult to gauge, modern scholarship having greatly modified original claims of vast human losses and near-total economic ruin. Nonetheless, in the most embattled realms, such as Württemberg, more than 50 percent of the people died or disappeared; elsewhere, the loss was less severe. Most historians agree that an overall population decline of 15 to 20 percent (from *c.* 20 million to 16 or 17 million) occurred during the war and the ensuing epidemics. In addition, historians agree that in theatres of war rural impoverishment and displacement of people were widespread, while economic regression happened nearly everywhere. For German society overall, the war was a traumatic experience; it is rivaled in the national consciousness only by the 1939–45 war as a time of unmitigated disaster.

To gain perspective on these calamities, their wider European aspects must be considered. Wars, uprisings, and political turmoil had occurred in many countries during the first half of the 17th century. The Fronde—a series of civil wars in France between 1648 and 1653 whose goal, at least in part, was to halt the growing power of royal government—and the Civil War in England (1642 to 1651) are only the most famous of these disturbances. Turmoil had occurred also in Catalonia, Portugal, Naples, Ireland, Scotland, Sweden, and Russia. Historians have referred to these events, including the numerous local manifestations of the Thirty Years' War, as parts of a general crisis in the fabric of European society, the causes of which range from a worsening of the climate (Little Ice Age) to plagues, often spread by the armies roaming Europe almost continuously at that time. But the most destabilizing factor burdening society was the centralizing monarchy with its expanding bureaucracy, extravagant courts, swollen armies, and incessant wars, all of them supported by heavy taxation. No social group, or estate, was unaffected by the effort of monarchs to alter in their favour traditional ways of distributing power and influence. Resentment of this and of its social cost was widespread; hence the proliferation and the scale of rebellions.

TERRITORIAL STATES IN THE AGE OF ABSOLUTISM

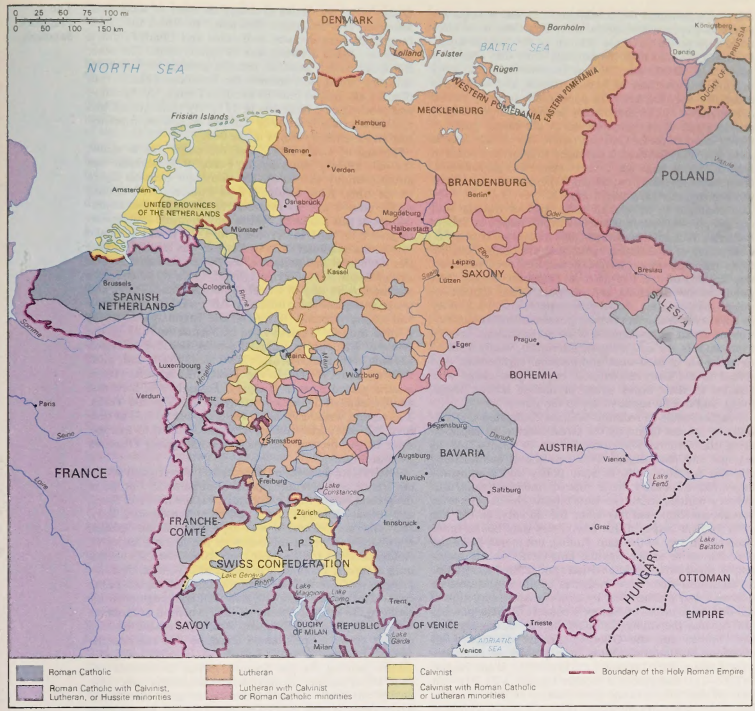
The empire after Westphalia. The empire was an awkward structure. German historians of an older, nationalistic generation deplored the fact that the empire lacked any of the attributes of a great power and lamented its victimization by more unified foreign states. Such critics always quote the 17th-century legal scholar Samuel Pufendorf, who called the empire a "monstrosity," and interpret this term as a value judgment rather than an expression indicating the inapplicability of standard categories of political classification. Recent scholars have been more appreciative of the post-1648 empire as a loose-jointed but not ineffective constitutional edifice within which could coexist 300 large and small, secular and ecclesiastical principalities, 51 imperial (*i.e.,* immediate to the empire) cities, and nearly 2,000 imperial counts and knights, each of whom possessed the same territorial sovereignty as an elector or a duke. The empire proved a working federation for the varied interests of these distinct sovereign entities, some of them large and powerful (like Saxony, whose electors were also kings of Poland from 1697 to 1763, or Brandenburg, whose prince was also king in Prussia) and some laughably tiny (such as the Abbey of Baindt in Swabia, a fully independent territory of a few hundred acres inhabited by 29 nuns and governed by a princess-abbess).



Germany in 1648.

The empire's administrative organs, especially the districts (*Kreise*), protected the small and weak from the predatory aims of the strong. Because most constituent members were vulnerable, there was no general inclination, despite disunity among the estates on matters of taxation and religious parity, to break the frame that guarded the status quo. The emperor's suzerainty over the entire realm went unchallenged, but virtually no real power adhered to his title, executive authority having been thoroughly particularized between 1555 and 1648. To prevent a resurgence of imperial power, princes formed alliances among themselves, such as the League of the Rhine (Rheinbund), tied in 1658 to France and Sweden. In the princely territories authority fell increasingly to the princes (exceptions were Württemberg and Mecklenburg), while territorial estates dwindled in political importance. In each of the empire's constituent units, estates served mainly to uphold established hierarchies and traditions, as did the empire as a whole. It was an inherently conservative system.

The consolidation of Brandenburg-Prussia and Austria. Against an overall tendency among the empire's constituent units to keep things as they were, the larger territories pursued an insistent policy of dynastic and personal aggrandizement. A number of factors favoured state building in the post-1618 era. General economic exhaustion made central direction of, and active intervention in, commerce and industry seem to be the only way out of stagnation. War taxes, raised to a steep level during the French wars of the 1670s (see below), greatly increased the financial might of rulers, who came to control an unprecedented share of society's wealth by preparing for and engaging in military conflict. Because territorial assemblies opposed this siphoning process—whose proceeds, augmented by subsidies from abroad, served mostly to create standing armies and a supporting state apparatus—rulers attempted to reduce even further the estates' role in policymaking. The nobility, growing economically dependent on princely service, adapted itself to an essentially ancillary function at court. In society at large the view gained ground that the country's welfare was safest with the ruler—a view vigorously promoted by official propaganda. Two of the empire's territories, Brandenburg-Prussia and Austria, profited above all others from these developments.



The range of confessions in Germany, 1650, as a result of the Thirty Years' War.

Frederick William, the "Great Elector"

The story of Brandenburg-Prussia has always been read as exemplifying the triumph of political skill and audacity over unfavourable conditions. Sparsely populated and deficient in resources, Brandenburg in 1648 was a patchwork of scattered territories. Its ruler, Frederick William (1640–88), later called for his achievements the "Great Elector," faced the problem of integrating and defending widely separated possessions, which included the duchy of Prussia, inherited in 1619 but remaining under Polish suzerainty and geographically separated from the electorate; the counties of Cleves, Mark, and Ravensberg in the Rhine and Westphalia regions, gained in 1614, also distant from Brandenburg and not contiguous with each other; and eastern Pomerania and various small lands and bishoprics acquired in the Treaty of 1648. Through nimble diplomatic maneuvering, such as changing sides several times between Sweden and Poland and beweeen France and the emperor, he augmented and solidified his realm and his authority within it: he won direct rule over Prussia as its duke and acquired the important episcopal territory of Magdeburg.

Frederick William's instrument in the attainment of these and subsequent prizes was the army, a permanent force of 30,000 disciplined professionals, the adequate financial support of which dictated every aspect of his government. Large revenues from taxes required a flourishing economy, the stimulation and direction of which by mercantilist principles was a main undertaking. Economic growth was further accelerated late in the Great Elector's reign by the influx of nearly 20,000 skilled Huguenot refugees following the revocation of the Edict of Nantes by Louis XIV in 1685 and by the resettlement of Dutch colonists. A territorywide system of state administration undergirded this economic and fiscal effort and resulted in the creation of a professional bureaucracy that permitted the Great Elector to govern essentially without estate participation. The land-owning nobility supported their prince in exchange for the freedom to exploit their peasants as they saw fit. In these ways Frederick William laid the foundation for what was to become an autocratically ruled state, enabled by its strong economy, tightly run administration, efficient fiscal organization, and powerful army to play a prominent role in the empire's and Europe's affairs.

"King in Prussia"

The Great Elector's efforts were rewarded in 1701 when his successor, Frederick III (1688–1713), obtained from the emperor (who needed the Brandenburg army for the impending War of the Spanish Succession) the right to style himself "King in Prussia." The title, recognized internationally upon the conclusion of the war in 1713, was of considerable importance to Brandenburg in its competition with Saxony, whose ruler had become king of Poland in 1697, for preeminence in north Germany. But it was Frederick's son, Frederick William I (1713–40; Prussian rulers renumbered themselves upon bestowal of the royal

title) who perfected the combination of statist structure, productive energy, and ethical drive that came to be identified with modern Prussia. Known as the "soldier-king," Frederick William built his standing army into a force of more than 80,000 men. It was the fourth largest army in Europe (after those of France, Russia, and Austria) in a country only thirteenth in population, superbly drilled and equipped but serving mainly defensive purposes. Only peasants and journeymen served in the ranks, many of whom were impressed abroad, while the middle classes were safe from the draft but obliged to quarter soldiers in their homes. A huge war chest obviated foreign subsidies, and reliable revenues, more than 70 percent of which went to the army, provided ample support.

To continue to draw high taxes without ruining land and people, the country's level of wealth had to be raised. Frederick William therefore pursued an aggressive mercantilist policy of stimulating agriculture and manufacturing while reducing unnecessary expenditures; even his court was stripped of many of its royal trappings. Export bans preserved raw materials, and sumptuary laws limited indulgence in luxuries. Town governments ceded authority to royal commissioners, whose powers included supervision of urban production. A work ethos suffused society from the top: the king's ascetic Calvinism, which dictated to him a life of hard work and personal engagement, was spread to his Lutheran subjects by a Pietist clergy who instilled in their flocks habits of intense labour, frugal living, and dutiful subservience to the state.

Organizationally, Frederick William completed the centralizing process begun by the Great Elector, its capstone being the General Directory, set up in 1723. Tied to regional and local organs by a network of commissioners, this supreme body of state policy and administration directed industry, trade, finance, internal affairs, and military matters in all the state's territories. Upper-level bureaucrats came entirely from the nobility, as did the army's officer corps; in this way nobles were bound more closely than ever to the state. Ruling, not merely reigning, over the entire edifice was the king-elector in his "cabinet," a small circle of close advisers and trusted secretaries. So successful were these measures in lifting the state to influence and prestige that by 1740 Prussia counted as a full-fledged member of Europe's concert of great powers.

In Austria, the ruling Habsburg house's lasting conflict with France and the Ottoman Empire dominated all questions of statecraft. With their powers as emperors greatly diminished, Leopold I (1658–1705), his son Joseph I (1705–11), and Joseph's brother Charles VI (1711–40) bent all their efforts to the consolidation of their dynastic and crown lands in central and eastern Europe. Although they failed to achieve Prussian-style streamlining, they raised Austria to the rank of a major state. The Habsburgs' conglomeration of territories included the Austrian duchies (Austria, Styria, Carinthia, Krain, and the county of Tirol), the Bohemian provinces (kingdom of Bohemia, Moravia, Silesia), the kingdom of Hungary, and—after 1714, following the War of the Spanish Succession—the southern Netherlands (Brabant, Luxembourg, Flanders) and the duchy of Milan.

These disparate lands were held together only by the Habsburg monarchy, but the monarchs were distracted from the task of integrating them. They were preoccupied by imperial concerns and by dynastic complications, notably the succession question. Until the reforms of Maria Theresa's reign (1740–80) Austrian administration never became effective. Finances were especially muddled, because tax administration remained with the estates of the various territories, along with control over other sources of revenue. The army of 100,000 men, though the third largest in Europe, was barely adequate for the defense of so large and scattered a realm. A supreme war council and a central financial chamber overlapped with special commissions created by the emperor's privy council, which also handled military and fiscal affairs. Nonetheless, the realm held together.

The prospect of a succession without a male heir, however, presented the severest test to the realm's cohesion. It became the chief enterprise of Charles VI to persuade the estates of his territories to accept an order of succession, known as the "Pragmatic Sanction," by which the Habsburg lands were declared indivisible and Charles's oldest daughter, Maria Theresa, was to inherit them. The other European powers assented, because splitting the Habsburg complex would have thrown the European balance into disarray and played into the hands of France; the Sanction was proclaimed a basic law in 1713. By then Austria had met successfully a series of Turkish incursions from the east and French invasions in the west.

The Pragmatic Sanction

The age of Louis XIV. For the empire as a whole, the half century following the Peace of Westphalia was almost entirely shaped by the dominant political figure of the time, King Louis XIV of France. The response of the empire and its members to the aggressive undertakings of this monarch, whose aim from his assumption of power in 1661 to his death in 1715 it was to make France the mightiest state in Europe, was largely reactive (for a different interpretation, see the article FRANCE: *The age of Louis XIV: Foreign affairs*). Only in its struggles against Louis's ally in the east, the Ottoman Turks, did the empire show some initiative. After a Polish relief army had helped imperial, Bavarian, and Saxon troops to lift a three months' Turkish siege of Vienna in 1683 in the Battle of Kahlenberg, imperial armies took the offensive, winning battles at Ofen (1686), Mohács (1687), and, most notably, Zenta (1697). In the Treaty of Carlowitz (1699) Austria gained parts of Hungary, Transylvania, Slavonia, and Croatia, all of them formerly occupied by the Turks. The eastern wars resumed in the early and mid-18th century, but the Turks were never again a threat to Europe, since Russia became the chief bulwark against Ottoman expansionism.

Matters were different on the empire's western and southern fronts. The overriding political question in Europe in the second half of the 17th century was the future of Spain and its vast holdings in the southern Netherlands, Italy, and America, because it was expected that the Spanish Habsburg line would die out with the feeble Charles II. Contenders for the Spanish inheritance were the Habsburg emperor, Leopold I, husband of a younger Spanish princess, and Louis XIV. The French monarch, son of the eldest daughter of Philip III, had further strengthened his claim to the Spanish throne in 1659 when, in accordance with the Peace of the Pyrenees, which had ended the long conflict between Spain and France, he had married the Spanish infanta.

While waiting for the Spanish throne to become vacant, Louis pursued an aggressive expansionist policy. He pushed his forces toward Germany to make the Rhine River France's new eastern border. In 1667 he occupied Flanders and in 1670 Lorraine; in 1672 he attacked Cleves and invaded the United Provinces of the Netherlands, his main antagonist in the wars that followed. In 1679 he began to penetrate Alsace, occupying the imperial city of Strassburg (Strasbourg) in 1681. Lacking the military power to bring the whole empire to its knees, Louis resorted to the lure of money; at one time or another almost every German state was in his pocket, either serving as ally or remaining neutral. Though not incapable of acting on national impulses, German princes—the Great Elector being a case in point—always served territorial interests first. This prevented the emperor, himself at times allied with Louis, from forging a solid front against France.

Leadership of the anti-France coalition passed to the Netherlands. William of Orange, as stadtholder of Holland and captain general of the United Provinces, emerged as the most determined opponent of French aggression. Upon becoming king of England in 1689, he changed the direction of English politics, which had been pro-French under the last Stuart king. The threat of a French universal monarchy arose dramatically as the death of the last Habsburg in Spain approached (Charles II died in 1700) and Louis's plans for a French claim on the entire Spanish inheritance swung into place. A Grand Alliance now formed against him (it was formally concluded in 1701), consisting of the empire (except Bavaria and the electorate of Cologne), the Netherlands, England, Sweden, Brandenburg-Prussia, and Savoy. Its aim was to restore

War of the Spanish Succession

the European balance to the status of 1648 and 1659 by ejecting Louis from his conquests and by splitting the Spanish empire.

The first phase of the ensuing struggle, known in Germany as the Palatine War, was fought in Germany. It led to savage destruction in the Palatinate and in Swabia but to no decisive victories on either side. A temporary peace (Rijswijk, 1697) forced Louis to make concessions and perhaps to realize the limits of his strength; his revocation of the Edict of Nantes in 1685 had set Protestants everywhere against him. The long-prepared War of the Spanish Succession broke out in 1701. The empire played a minor role in this conflict, Louis having failed in his diplomatic effort to enlarge the simultaneous Northern War of Sweden against Denmark, Russia, and Saxony into an attack on Brandenburg-Prussia. Austria, however, proved a major antagonist in this last attempt by the *grand monarque* to gain the upper hand. While Austrian and English troops operated in Italy, French and Bavarian forces fought in southern Germany. Neither side won clear victories, but the alliance gradually gained the advantage, until the death of emperor Joseph I in 1711 placed his brother Charles, who had been proclaimed Spanish king, on the imperial throne as Charles VI (1711–40). This raised the spectre of a Habsburg reunion of the Holy Roman and Spanish empires, which was no more agreeable to European powers than the prospect of French overlordship. Thus the alliance severed.

Peace negotiations began in 1712, resulting in a number of treaties, signed at Utrecht and Rastatt in 1713–14. The Spanish empire was partitioned, with the Spanish Netherlands, Milan, Naples, and Sicily going to Austria and Spain itself coming under the rule of Philip V of Bourbon, a grandson of Louis XIV. The alliance's original aim, to prevent French hegemony, had been achieved, though in the follow-up War of the Polish Succession (1733–35) France acquired control of Lorraine. Austria profited substantially in territorial terms, and a few other German rulers profited as well, albeit less so. As for the empire itself, it had gained no real benefits from more than half a century of intermittent warfare.

German society, however, was deeply affected. Economic stagnation and slow demographic recovery after the Thirty Years' War made Germany dependent on governmental intervention as a means of stimulating recovery. This left the country exposed to foreign influences, which reached land and people by way of the many princely courts and the elites clustered there.

German cultural life took its cues from abroad: Baroque art—the preeminent expression of monarchical power and of Roman Catholic resurgence after the Reformation—came from Spain and Italy, opera from Italy, and polite language and manners from France. The style of this period took French patterns as its model, from elaborately coded court ceremonials to dress, social conventions, food, and conversation. French absolutism not only became the political model, however scaled down, for the governance of all states in the empire, but every German prince and princeling imitated the lavish display with which Louis XIV created his aura of majesty and outshone his rivals. This started up a lively domestic market in luxuries, not to mention splendid works of architecture and decoration. But the cost of these luxuries was prohibitive (in 1719 the Palatine court consumed 50 percent of the territory's revenues) and represented an enormous burden on the people, especially when added to the cost of large armies and proliferating bureaucracies. Not only did this conspicuous consumption widen the social division between the court-oriented elite and the bulk of the urban and rural population but the culture's foreign provenance of the goods also inhibited creative impulses at home.

In the second half of the 17th century German energies were to a large extent still focused on religion. The confessional pluralism legitimized by the settlement of 1648 encouraged emphasis on theological distinctions, exacerbating the move toward religious orthodoxy under way in each denomination since the 16th century. The one genuinely German product of this religious preoccupation was Pietism. It was a religious movement within Lutheranism that opposed rigid dogmatism and promoted instead a subjective, mystical devoutness and an emphasis on a pious life guided by love of one's neighbour as well as of God. Influenced by English Puritanism, Pietism was shaped in its theology by Philipp Jakob Spener (1635–1705) and in its organization by his disciple August Hermann Francke (1663–1727), who established a centre for its promulgation in the Brandenburg city of Halle. There he founded schools, orphanages, medical facilities, and a printing house for publishing cheap Bibles and devotional works, which made of Pietism a widely influential program of evangelical activism. The intensely emotional and mystical flavour of Pietist poetry is preserved in the cantata texts set to music by Johann Sebastian Bach, in whose deeply spiritual church music Protestant chorale singing, another indigenous German product, reached its apogee.

The contest between Prussia and Austria. In 1740 the death of the Habsburg emperor Charles VI without male heir unleashed the most embittered conflict in Germany since the wars of Louis XIV. The question of who would ascend the Austrian throne had occupied statesmen for decades. Rival claimants disputed the right—by the terms of the Pragmatic Sanction (1713)—of Charles's daughter Maria Theresa to succeed; France supported them, its aim being, as before, the fragmentation of the Austrian state. But it was the new Prussian king, Frederick II (1740–86), who began the conflict. To understand what follows, the modern reader should remember that few observers, even in the enlightened 18th century, disputed a ruler's right to do what he wished with his state. Dynastic aggrandizement, territorial expansion, prestige, honour, power, and princely glory were legitimate grounds for war and sound reasons for demanding the sacrifices necessary to wage it. The only position from which to oppose this arrogation was the Christian ethic, but to do so had proved futile when last tried by Erasmus and Sir Thomas More in the 16th century. No checks—philosophical, moral, or political—therefore restrained kings from indulging their taste for conquests.

Soon after taking control of affairs, Frederick reversed his father's cautious policy of building and hoarding, rather than consuming, Brandenburg-Prussia's military potential. He attacked Silesia, a province in the kingdom of Bohemia and thus part of the Habsburg monarchy, which Prussia had long desired for its populousness, mineral resources, and advanced economy. In exchange for an Austrian cession of Silesia he offered to accept the Pragmatic Sanction (formally recognized by his predecessor in the Treaty of Berlin in 1728) and support the candidacy of Maria Theresa's husband, Francis Stephan, as emperor. But the resolute woman who now headed the Austrian Habsburgs (1740–80) decided to defend the integrity of her realm, and war began in 1740. Austria was helped only by a Hungarian army, though financial support came from Austria's ally, England. Prussia was joined by Bavaria and Saxony in the empire as well as by France and Spain. The Prussian armies, though greatly outnumbered against forces that included troops from Russia and Sweden, revealed themselves as by far the best as well as the best led. In the Treaty of Hubertusburg of 1763 Prussia kept Silesia but could not hold on to Saxony, which Frederick had invaded in 1756 and which he coveted as a way of rounding out his state.

War of the Austrian Succession

In a sense, the War of the Austrian Succession (1740–48) was another of the many internal struggles over the constitutional balance in the empire in which territorial states opposed imperial authority. But it was also part of an international struggle, with France and England fighting out their rivalry in western and southern Europe, America, and India. Thus it blended into the worldwide Seven Years' War (1756–63), the latter being preceded by a "Diplomatic Revolution" in which England switched its support to Prussia and France allied itself with its traditional foe, Austria (a part of this agreement was the marriage, in 1770, of the Austrian princess Marie Antoinette to the future Louis XVI). The real significance of the Seven Years' War lay in the Peace of Paris of 1763, which concluded for a time the maritime and colonial conflict between France and England.

Austria having survived the war and Prussia having increased in size and immeasurably in prestige, these two powers now dominated German affairs in a condition of tension usually called "the German dualism," meaning that each had become so powerful that only the other could keep it in some sort of check. The monarchs of both realms carried out important internal reforms. Guided by her interior minister, Count Friederich Wilhelm Haugwitz, Maria Theresa streamlined the Austrian administrative structure on the Prussian model, thus drawing together, to the extent possible, the multiethnic and polyglot regions of the far-flung Habsburg empire. The remaining powers of the estates were curtailed everywhere and centralization institutionalized in absolutist fashion but without attaining the full integration of the Prussian system. Maria Theresa's son, Joseph II (1765–90), completed this program of modernization.

In Prussia, Frederick II further tightened his control of all aspects of public life in his far-flung kingdom. But in accordance with his personal commitment to rational tolerance and free-thinking skepticism he also undertook extensive legal reforms; he virtually abolished judicial torture, lifted some of the tax burden from the poorest of his subjects, established religious tolerance as a policy of his state, and encouraged scientific and scholarly activity in the Prussian Academy. Like his father, he was a vigorous promoter of economic development and colonization. His taste for French Enlightenment thought and his own prolific creativity in letters and music lent to his reign the flavour of an era shaped by a philosopher-king, albeit one with the instincts of a ruthless power politician. His successes in war and peace earned him a place as national hero as well as the title "the Great." (Ge.St.)

Germany from *c.* 1760 to 1871

GERMANY TO 1815

Germany in the middle of the 18th century was a country that had been drifting in the backwaters of European politics for more than a hundred years. The decisive roles in the affairs of the Continent were played by those great powers—such as France, England, and Spain—whose economic resources and commercial connections provided a solid foundation for their military might. The states of central Europe, on the other hand, floundered in a morass of provincialism and particularism. All the forces that had contributed to the rise of powerful national monarchies west of the Rhine were lacking in the east. In the Holy Roman Empire the central government was losing rather than gaining strength, the princes were enlarging their authority at the expense of the crown, and business initiative was being discouraged by the lack of political unity and by the remoteness of the major trade routes.

The Holy Roman Empire

Civic power became increasingly concentrated in the hands of local governments controlled by aristocratic overlords, ecclesiastical dignitaries, or municipal oligarchs. The history of Germany between the Thirty Years' War and the French Revolution is largely the sum total of the histories of dozens upon dozens of small political units, each enjoying virtually full rights of sovereignty. The rulers of these gingerbread principalities, copying the example of the royal court of France or Austria, built costly imitations of the palaces of Versailles and Schönbrunn, which today are the delight of tourists but which were once the curse of an impoverished peasantry. The tradition of princely authority, an instrument of national greatness in western Europe, encouraged national divisiveness in central Europe. The petty rulers of Germany legislated at will, levied taxes, concluded alliances, and waged wars against each other and against the emperor. Policies pursued in Munich, Stuttgart, Dresden, or Darmstadt reflected policies originating in Paris, Vienna, London, or Madrid but without seeking a goal greater than the promotion of particularistic interests.

Political institutions designed theoretically to express the will of the nation continued to function, yet they had become an empty shell. The Holy Roman emperor was still elected in accordance with a time-honoured ritual that proclaimed him to be the successor of Caesar and Augustus. The splendid coronation ceremony in Frankfurt am Main, however, could not disguise the fact that the office conferred on its holder little more than prestige. Since all the emperors except Charles VII were Habsburgs by birth or marriage, they enjoyed an authority that had to be respected. But that authority rested not on the prerogative of the imperial crown but on the possession of hereditary lands stretching from Antwerp in the west to Debrecen in the east. The sovereigns of the Holy Roman Empire, in other words, were able to play an important role in German affairs by virtue of their non-German resources. And, since Germany was not the main source of their strength, Germany was not the main object of their concern. The emperors tended to regard the dignity bestowed upon them as a means of furthering the interests of their dynastic holdings. The Imperial Diet meeting in Regensburg had also become an instrument for the promotion of particularistic advantage rather than national welfare. It continued in theory to express the will of the estates of the realm meeting in solemn deliberation. In fact it had degenerated into a debating society without authority or influence. The princes had ceased to attend the sessions, so that only diplomatic representatives were left to discuss questions for which they were powerless to provide answers. The other central institutions of the empire, such as the imperial cameral tribunal in Wetzlar, languished in indolence. Constitutionally and politically, Germany in about 1760 resembled Poland in that a once vigorous and proud state had become weakened by internal conflict to the point that it invited the intervention of its more powerful neighbours.

What saved Germany from the fate of Poland was the ability of one of the member states to defend the empire against aggression. For 200 years Austria acted as the bulwark of central Europe against French expansion. Its possessions, forming a chain of protective bases extending between the North Sea and the Danube, had time and again borne the brunt of attacks by Bourbon armies. The frontiers of France kept moving closer to the Rhine, but the Holy Roman Empire was at least spared the tragedy of partition that befell the Polish state. It was partly in recognition of the vital role that the Habsburgs played in the defense of Germany that the electors chose them as emperors with such regularity. The Austrian monarchy, moreover, endowed with resources comparable to those of the western nations, was able to pursue a policy of political rationalization with greater success than the principalities. The rulers in Vienna succeeded in improving the administration, strengthening the economy, and centralizing the government. Until the middle of the 18th century, Austria remained the only great power east of the Rhine.

The traditional role of the Austrian Habsburgs

Further rise of Prussia and the Hohenzollerns. The emergence of the Hohenzollerns as rivals of the Habsburgs and the beginning of the Austro-Prussian dualism created the possibility of reversing the process of civic decentralization that had prevailed in central Europe since the late Middle Ages. The interests of the territorial princes of the Holy Roman Empire inclined them toward a policy of particularism, while the government of Austria, with its Flemish, Italian, Slavic, and Magyar territories, could not perforce become the instrument of German unification. Prussia, on the other hand, was militarily strong enough and ethnically homogeneous enough to make national consolidation the main object of statecraft. But, though the creation of a united Germany became its mission, this mission was not from the outset one that it accepted willingly or even consciously. The intention of Frederick the Great and of his successors Frederick William II and Frederick William III was to pursue dynastic rather than national objectives. Like the lesser princes of central Europe, all they sought was to maintain and enlarge their authority against the claim of imperial supremacy. Far from wanting to end the disunity of Germany, they hoped to prolong and exploit it. The patriotic Prussophile historians, who a hundred years later argued that what Bismarck had achieved was the consummation of what Frederick had sought, were letting the present distort their understanding of the past. In fact, the greatest of the Hohenzollerns had remained as indifferent to the glaring

political weaknesses of his nation as to its great cultural achievements. His attitude toward the constitutional system of the Holy Roman Empire was similar to that of the self-seeking princelings who were his neighbours and from whom he was distinguishable only by talent and power. He may have scorned their sybaritic way of life, but politically he wanted what they wanted—namely, the freedom to seek the advantage of his dynasty without regard for the interests of Germany as a whole.

His preoccupation with the welfare of his state rather than with that of his nation is apparent in the strategy by which he tried to check Habsburg ambitions after the Seven Years' War (1756–63). During the first half of his reign he had relied primarily on military force to restrict and undermine imperial authority. In the second half he preferred to employ the weapons of diplomacy to achieve the same end. In 1777 the ruling dynasty of Bavaria came to an end with the death of Maximilian Joseph. The elector of the Palatinate, Charles Theodore, now became ruler over the territories of both branches of the house of Wittelsbach. Without legitimate offspring to whom to leave his state and without affection for his newly acquired eastern possessions, he agreed to a plan proposed by Emperor Joseph II according to which part of the Bavarian lands would be ceded to Austria. But any increase in the strength of the Habsburgs was unacceptable to Frederick the Great. With the tacit approval of most of the princes of the empire, he declared war against Austria in 1778, hoping that other states within and outside central Europe would join him. In this expectation he was disappointed. Yet Joseph also became discouraged by the difficulties encountered in what he had believed would be an easy success. The War of the Bavarian Succession dragged on from the summer of 1778 to the spring of 1779, with neither side enhancing its reputation for prowess on the field of battle. There was much marching back and forth, while hungry soldiers scrounged for food in what came to be called the "Potato War." The upshot was the Treaty of Teschen (May 1779), by which the Austrian government abandoned all claims to Bavarian territory except for a small strip along the Inn River. The conflict had brought Frederick no significant military victories, but he had succeeded in frustrating Habsburg ambition.

The War of the Bavarian Succession

Joseph II, however, was a stubborn adversary. In 1785 he once again advanced a plan for the acquisition of Wittelsbach lands, this time on an even more ambitious scale. He suggested to Charles Theodore nothing less than an outright exchange of the Austrian Netherlands for all of Bavaria. The emperor, in other words, proposed to surrender his distant possessions on the North Sea, which were difficult to defend, for a territory that was contiguous and a population that was assimilable. The scheme went far beyond that which Prussia had defeated seven years before, and Frederick opposed it with equal determination. He hoped to enlist the diplomatic aid of France and Russia against what he regarded as an attempt to upset the balance of power in central Europe. But, more than that, he succeeded in forming an Association of Princes, which 17 of the more important rulers in Germany joined. The members pledged themselves to maintain the fundamental law of the empire and to defend the possessions of the governments included within its boundaries. The growing opposition to the absorption of Bavaria by Austria persuaded Joseph that the risks inherent in his plan outweighed its advantages. The proposed exchange of territories was dropped, and Frederick could celebrate yet another triumph of his statecraft, the last of an illustrious career. But the association of princes that he founded did not survive its author. Its sole purpose had been the protection of princely prerogative against imperial authority. Once the danger had passed, it lost the only justification for its existence. Those nationalists who later maintained that it foreshadowed the creation of the German Empire misunderstood its origins and objectives. It was never more than a weapon in the struggle for the preservation of a decentralized form of government in central Europe.

The subordination by the Hohenzollerns of national to dynastic interests was even more apparent in the role they played during the partitions of Poland. Frederick the Great was the chief architect of the first partition, that of 1772, by which the ill-starred kingdom lost about a fifth of its inhabitants and a fourth of its territory to Prussia, Russia, and Austria. His successor, Frederick William II, helped to complete the destruction of the Polish state by the partitions of 1793 (between Prussia and Russia) and 1795 (between Prussia, Russia, and Austria). The result was bound to be an enhancement of Prussia's role in Europe but also a diminution of its role in Germany. The Hohenzollerns willingly embarked on a course that would in time have transformed their kingdom into a binational state comparable to the Habsburg empire. The German population in the old provinces would have been counterbalanced by the Slavic population in the new; the Protestant faith of the Prussians would have had to share its influence with the Roman Catholicism of the Poles; the capital city of Berlin would have found a competitor in the capital city of Warsaw. In short, the centre of gravity of the state would have shifted eastward, away from the problems and interests of the Holy Roman Empire. Yet the rulers of Prussia did not shrink from a policy that was likely to have such far-reaching consequences. They never contemplated sacrificing the advantage that their state would gain from an enlargement of its resources in order to assume the role of unifiers of their nation. Such a political attitude would have been an anachronism during the age of princely absolutism in central Europe. It was not design but accident that before long led to the abandonment by Prussia of most of its Polish possessions and that thereby allowed it to continue to play a vital part in the affairs of Germany.

Contingent results of the Polish partition

The cultural scene. Whereas in England the great literary epoch of Queen Elizabeth I had coincided with commercial and naval expansion, and in France the golden age of classicism had added lustre to the military glory of Louis XIV, German arts and letters flourished amid tiny principalities and somnolent towns that could only envy the powerful national monarchies west of the Rhine. Moreover, whereas in France and England, where public opinion could exert a significant influence on government, the debate over issues of state and society was conducted with a vigour that reflected its importance, in Germany the debate was bound to remain purely theoretical. No Voltaires, Rousseaus, or Burkes were likely to emerge out of such an environment. The thinkers of central Europe tended to emphasize introspection and spirituality. Culture became an escape from the narrow world of princely absolutism. Intellectual energies that could not reform the community fought to emancipate the individual through self-purification and self-perfection.

This was the background of German idealism, a philosophical movement seeking to liberate ethics and aesthetics from the shackles of empirical knowledge. Armed with the weapons of Kantian thought, it attempted to prove that there was a realm of experience lying beyond the categories of scientific investigation: the realm of the good, the true, and the beautiful. There were realities of the spirit and the mind, in other words, that were inaccessible to the practicality of the British empiricists or the intellectualism of the French rationalists. The disciples of idealism hoped to transcend the barriers created by nation, class, and religion. They spoke in the name of humanity as a whole, which manifested its underlying harmony through the infinite variety of its political, social, and theological categories. Gotthold Ephraim Lessing pleaded for religious toleration on the basis of a common system of ethical values to which all men of goodwill could subscribe. Johann Gottfried Herder preached that the unique character and meaning of each culture contributed to the richness of common humanity that defied state boundaries. Johann Joachim Winckelmann deified the classical ideal of beauty that he found in Greek art as an eternal standard, immune to the vicissitudes of time and history. These were views that offered an escape from the narrowness of everyday life. Men who found no scope for their talents in the petty world of princely authority could turn to the liberating spirituality of the idealist philosophy. Thought in central Europe gradually acquired a metaphysical coloration that distinguished it from the more robust pragmatism of phi-

German idealism

losophy in the west. It was during the second half of the 18th century that the Germans began to consider their country "the land of thinkers and poets."

The literary revival of the age displayed the same quality of introspective idealism as the philosophical movement. Johann Wolfgang von Goethe, the greatest genius of German letters, willingly accepted the existing system of civic and social values. He regarded the disunity of his nation as an expression of its historic character and defended the authority of the petty princes as an instrument of good government. He urged his countrymen to seek greatness not in collective action but in individual perfectibility. After a period of youthful rebellion against traditional canons of literary propriety, he turned to a classicism in which a serene acceptance of life harmonized with his own sympathy for the established order. Friedrich Schiller, a man of more turbulent temperament, felt a sense of resentment against political injustice and weakness. In his plays and poems there are occasional outbursts of indignation and appeals for reform. Yet there is also a pessimistic mood of resignation induced by the burden of civic ineffectualness that history had imposed on his people. Ultimately, he too sought refuge from the world in the poet's private vision. The Sturm und Drang ("Storm and Stress") was a movement of literary innovation through which a group of young writers in the last decades of the 18th century sought to throw off the yoke of accepted standards of composition, but it remained confined to problems of prosody and taste, refusing to grapple with political or social issues.

The Sturm und Drang movement

The cultural achievements could not alter the harsh realities of national fragmentation and princely autocracy. They supported, however, the ideals of rational reform and social progress that the Enlightenment had introduced throughout the Continent. In Germany as elsewhere the 18th century became the age when the monarchical principle advanced the loftiest justification of its claim to power. The authority of the prince, so the argument went, was to be exercised not for his private advantage or gratification but for the greatness of his state and the welfare of his people. His power had to be unrestricted so that his benevolence might be unlimited. Absolute government was the only effective instrument for achieving the general good. Impressed by the scientific discoveries and material advances that they saw about them, men began to believe that the prejudices and injustices that had plagued society would gradually disappear before the steady march of reason.

Enlightened reform and benevolent despotism. The main source of enlightened reform was to be the crown, but many well-intentioned people of means and education also began to apply a new standard of conduct in their dealings with their fellow man. This change in attitude was apparent in the decline of religious resentments and discriminations. Never before had the relationship between Roman Catholics and Protestants among the well-to-do classes of central Europe been as free of rancor as on the eve of the French Revolution. It was at this time also that the Jews first began to emerge from the isolation to which a deep-seated intolerance had consigned them. The idea of assimilation held out to them the prospect of escape from the ghetto on the condition that they identify themselves in thought, speech, and attitude with the Christian society in which they lived. That prospect was to attract the Jewish minority in Germany more and more during the next 150 years. Religious toleration, however, was not the only article of faith of the Enlightenment. Its vision of a happier future included the reformation of education, the abolition of poverty, the alleviation of sickness, and the elimination of injustice. Men of goodwill established schools, founded orphanages, built hospitals, improved farming methods, modernized industrial techniques, and tried to raise the standard of living of the masses. While the hopes of the enlightened reformers of the 18th century far outstripped their accomplishments, the practical results of their efforts should not be underestimated.

According to the doctrines of benevolent despotism, however, the chief instrumentality for the improvement of society was not private philanthropy but government action. The state had the primary responsibility for preparing the way for that golden age which, in the opinion of many intellectuals, awaited humankind. The extent to which official policy conformed to rationalist theory depended, in central Europe as elsewhere, on the personality and ability of the ruler. Both of the leading powers of the Holy Roman Empire followed the teachings of benevolent despotism but with substantially different results. The emperor Joseph II, a well-meaning though doctrinaire reformer, attempted to initiate a revolution from above against the opposition of powerful forces that continued to cling to tradition. In the course of a single decade he tried to centralize the government of his far-flung domains, reduce the influence of the church, introduce religious toleration, and ease the burden of serfdom. His uncompromising program of innovation, however, alienated the landed aristocracy, whose support was essential for the effective operation of the government. The emperor encountered mounting unrest that did not end until his death in 1790 and the subsequent abandonment of most of the reforms that he had promulgated. Frederick the Great was more successful as an enlightened autocrat, but only because he was more cautious. His reorganization of the government was not as drastic, his belief in religious toleration remained less profound, and his assistance to the peasants did not go beyond a prohibition against the absorption of their holdings by the nobility. He invited settlers to cultivate reclaimed lands, and he encouraged entrepreneurs to increase the industrial capacity of Prussia. Among his most important accomplishments, although it was not completed until after his death, was the Prussian Legal Code, which defined the principles and practices of an absolute government and a corporative society. Yet Frederick was also convinced that the Prussian landed noblemen, the Junkers, were the backbone of the state, and he continued accordingly to uphold the alliance between crown and aristocracy on which his kingdom had been built.

Attempts at social and political reform

The achievements of benevolent despotism among the minor states of the Holy Roman Empire varied considerably. Some princes employed their inherited authority in a serious effort to improve the lot of their subjects. Charles Frederick of Baden, for example, devoted himself to the improvement of education in his margravate, and he even abolished serfdom, though without eliminating manorial obligations. Charles Augustus of Saxe-Weimar-Eisenach was a hardworking administrator of his small Thuringian principality, whose capital he transformed into the cultural centre of Germany. Charles Eugene of Württemberg, on the other hand, led a life of profligacy and licentiousness in defiance of protests by the estates of the duchy. Frederick II of Hesse-Kassel was another princely prodigal whose love of pleasure impoverished his subjects and forced his soldiers into mercenary service for England. The record of enlightened autocracy in central Europe was as uneven as in western Europe. Yet the ideas of the Enlightenment even at their best were unable to transform the basis of political life in the Holy Roman Empire. They could palliate, reform, and improve, but they could not alter a system of particularistic sovereignty and absolutistic authority resting on a hierarchical structure of society. They could not become an instrument of national consolidation or representative government. Only some great creative disruption of existing civic institutions could break through the crust of habit and tradition sanctified by history. Germany lacked the vital energies required for a process of political reconstruction. The galvanizing forces of rejuvenation and regeneration were to come from the outside.

The French Revolutionary and Napoleonic era. The French Revolution, transforming the Bourbon kingdom into a constitutional state, aroused intense excitement east of the Rhine. Most German intellectuals were at first in sympathy with the new order in France, hoping that the defeat of royal absolutism in western Europe would lead to its decline in central Europe as well. The princes, on the other hand, were from the outset fearful of the Revolution, which they regarded as a serious danger, for the example of unpunished insubordination by the French might encourage demands for reform among the Germans. The result was a growing hostility between the government in Paris and the rulers of the Holy Roman

The War of the First Coalition

Empire, which led in the spring of 1792 to the outbreak of war (the War of the First Coalition, 1792–97). The immediate occasion of the conflict was a quarrel over the rights of German princes with holdings in France and over the propagandistic activities of French émigrés in Germany. But the underlying cause was the clash of two incompatible principles of authority divided by profound differences regarding the nature of political and social justice. The course of hostilities soon revealed that the civic ideals and military tactics of the French Revolution were more than a match for the decrepit Holy Roman Empire. After 1793 the left bank of the Rhine remained under the control of France, and for the next 20 years its inhabitants were governed from Paris. Yet there is no evidence that they were dissatisfied with French rule or at least that they strongly opposed it. Devoid of a sense of nationalism and accustomed to submission to authority, they accepted their new status with the same equanimity with which they regarded a succession to the throne or a change in the dynasty. The Prussians, moreover, discouraged by defeats in the west and eager for Polish spoils in the east, concluded a separate peace at Basel in 1795 by which they in effect recognized the French acquisition of the Rhineland. The Austrians held out two years longer, but the brilliant successes of the young Napoleon Bonaparte forced them to accept the loss of the left bank in the Treaty of Campo Formio (Oct. 17, 1797).

End of the Holy Roman Empire. The peace proved short-lived, however, for at the end of 1798 a new coalition directed against France was formed (the War of the Second Coalition, 1798–1802). This time Prussia remained neutral. Frederick William III, a conscientious and modest but ineffectual ruler, was distinguishable from his father by private morality rather than political skill. The government in Berlin drifted back and forth, dabbling in minor economic and administrative reforms without introducing a significant improvement in the structure of the state. A decade of neutrality was frittered away while the army commanders rested on the laurels of Frederick the Great. Austria, on the other hand, played the same leading role in the War of the Second Coalition as in the War of the First Coalition, with the same unfortunate result. The French victories at Marengo (June 14, 1800) and Hohenlinden (Dec. 3, 1800) forced Emperor Francis II to agree to the Treaty of Lunéville (Feb. 9, 1801), which confirmed the cession of the Rhineland. More than that, those rulers who lost their possessions on the left bank under the terms of the peace were to receive compensation elsewhere in the empire. In order to carry out this redistribution of territory, the Imperial Diet entrusted a committee of princes, the Reichsdeputation, with the task of drawing a new map of central Europe. The major influence over its deliberations, however, was exercised by France. Napoleon had resolved to utilize the settlement of territorial claims to achieve a fundamental alteration in the structure of the Holy Roman Empire. The result was that the Final Recess (Hauptschluss) of the Reichsdeputation of February 1803 marked the end of the old order in Germany. In their attempt to establish a chain of satellite states east of the Rhine, the French diplomats brought about the elimination of the smallest and least viable of the political components of central Europe. They thereby also furthered the process of national consolidation, since the fragmentation of civic authority in the empire had been a mainstay of particularism. That it was not Napoleon's intention to encourage unity among his neighbours goes without saying. Yet he unwittingly prepared the way for a program of centralization in Germany that helped to frustrate his own plans for the future aggrandizement of France.

The end of the old order

The chief victims of the Final Recess were the free cities and the ecclesiastical territories. They fell by the dozens. Too weak to be useful allies of Napoleon, they were destroyed by the ambition of their French conquerors and by the greed of their German neighbours. They could still boast of their ancient history as sovereign members of the Holy Roman Empire, but their continued existence had become incompatible with the establishment of effective government in central Europe. The principal heirs to their holdings were the larger secondary states. To be sure, Napoleon could not keep Austria and Prussia from making some gains in the general scramble for territory that they had helped make possible. But he was especially solicitous for the welfare of those German rulers, most of them in the south, who were strong enough to be valuable vassals but not strong enough to be potential threats. Bavaria, Württemberg, Baden, Hesse-Darmstadt, and Nassau were the big winners in the competition for booty that had been the main object of the negotiations. Napoleon's strategy had been in the classic tradition of French diplomacy, the tradition of Richelieu and Mazarin. The princes had been pitted against the emperor to enhance the role that Paris could play in the affairs of central Europe. Yet neither the states that gained nor those that lost territory felt resentment at being used as pawns in a political game to promote the interests of a foreign power. Whatever objections they raised against the settlement of 1803 were based on expediency and opportunism. The most serious indictment of the old order was that in the hour of its imminent collapse none of the rulers attempted to defend it in the name of the general welfare of Germany.

The Final Recess was the next to the last act in the fall of the Holy Roman Empire. The end came three years later. In 1805 Austria joined the third coalition of Great Powers determined to reduce the preponderance of France (the War of the Third Coalition, 1805–07). The outcome was even more disastrous than its participation in the first and the second coalitions. Napoleon forced the main Habsburg army in Germany to surrender at Ulm (Oct. 17, 1805); then he descended on Vienna, occupying the proud capital of his enemy; and finally he inflicted a crushing defeat on the combined Russian and Austrian armies at Austerlitz in Moravia (Dec. 2, 1805). Before the year was out Francis II had been forced to sign the humiliating Treaty of Pressburg (December 26), which signified the end of the dominant role his dynasty had played in the affairs of central Europe. He had to surrender his possessions in western Germany to Württemberg and Baden, and the province of Tirol to Bavaria. Napoleon's strategy of playing off princely against imperial ambitions had proved a brilliant success. The rulers of the secondary states in the south had supported him in the war against Austria, and in the peace that ensued they were richly rewarded. Not only did they share in the booty seized from the Habsburgs but they also were permitted to absorb the remaining free cities, petty principalities, and ecclesiastical territories. Finally, in an assertion of the rights of full sovereignty, the rulers of Bavaria and Württemberg assumed the title of king, while the rulers of Baden and Hesse-Darmstadt contented themselves with the more modest rank of grand duke. The last vestiges of the imperial constitution had now been destroyed, and central Europe was ready to receive a new form of political organization reflecting the power relationship created by the force of arms.

The Confederation of the Rhine

In the summer of 1806, 16 of the secondary states, encouraged and prodded by Paris, announced that they were forming a separate association to be known as the Confederation of the Rhine. Archbishop Karl Theodor von Dalberg was to preside over the new union as the "prince primate," while future deliberations among the members were to establish a college of kings and a college of princes as common legislative bodies. There was even talk of a "fundamental statute" that would serve as the constitution of a rejuvenated Germany. Yet all these brave plans were never more than a facade for the harsh reality of alien hegemony in central Europe. Napoleon was proclaimed the "protector" of the Confederation of the Rhine, and a permanent alliance between the member states and the French Empire obliged the former to maintain substantial military forces for the purpose of mutual defense. There could be no doubt whose interests these troops would serve. The secondary rulers of Germany were expected to pay a handsome tribute to Paris for their newly acquired sham sovereignty. On August 1 the confederated states proclaimed their secession from the empire, and a week later, on Aug. 6, 1806, Francis II announced that he was laying down the imperial crown. The Holy Roman Empire thus came officially to an end after a history of a thousand years.

Period of French hegemony in Germany. The long conflict between emperors and princes in Germany had concluded with the triumph of the latter. Yet the victors soon discovered that, instead of achieving independence, they had merely exchanged one master for another. Indeed, they were more subordinate now to the wishes of Paris than they had been to those of Vienna. Napoleon gradually induced or forced all the states of Germany except Austria and Prussia, 36 states in all, to join the confederation. The inclusion of the two leading powers of central Europe might have proved troublesome for him or even dangerous. The participation of the secondary states, on the other hand, provided him with reliable mercenaries who owed him too much and feared him too much to oppose his wishes. He was free to do as he liked in the region between the Rhine and the Elbe. In order to enforce his embargo on continental trade with England, he annexed the entire coastline along the North Sea. When that was not enough, he added Lübeck on the Baltic to the French Empire. He carved out the kingdom of Westphalia for his brother Jérôme and the grand duchy of Berg for his brother-in-law Joachim Murat. He was the undisputed master of all of middle Germany.

After the formation of the Confederation of the Rhine, there was only one state in central Europe that had not yet been forced to submit to France. But the leaders of Prussia hesitated and wavered in their policy until they lost the opportunity of profiting from the War of the Third Coalition. Had they joined Austria and Russia against Napoleon, they might have kept him from gaining hegemony over Germany. Or had they become the allies of Napoleon, they might have established a sphere of influence in the region north of the Main River. As it was, they waited until they fell between two stools. They finally declared war against the French in October 1806, after Austria had been forced to surrender, Russia had decided to retreat, and the secondary states had become the vassals of Paris. Yet public opinion in the Prussian capital remained confident that the army of Frederick the Great would prove a match for the conqueror of Europe.

The subjection of Prussia

The result of such self-deception was a military disaster of unparalleled magnitude. In the two simultaneous battles of Jena and Auerstädt (Oct. 14, 1806) the Hohenzollern armies were completely routed, and the road to Berlin lay open before the French invaders. The city was occupied on October 27.

More disastrous than the military defeat, however, was the moral collapse of a state that had taught its citizens that obedience to authority was the supreme political virtue. The civilian population never thought of offering resistance to the advancing enemy. Even many army officers were so disheartened by Napoleon's success that they surrendered one fortified position after another without a fight. Frederick William III had to pay a terrible price for the policy of his ancestors, who had built efficient government at the expense of civic initiative. He tried to hold out in East Prussia, hoping that the Russian armies, which were still at war with Napoleon, would help him regain the rest of his kingdom. But when in July 1807 Alexander I concluded peace with France at Tilsit, the unfortunate Hohenzollern had no choice but to follow suit. The treaty that he was forced to sign was a catastrophe. Prussia lost almost half its territory and population, including most of the Polish possessions in the east as well as all of the territories west of the Elbe. Subsequent agreements, moreover, imposed a heavy indemnity, a military occupation, and a reduction in the size of the army. The proud monarchy of Frederick the Great had been reduced to a secondary state in Germany.

Central Europe remained under the dominant influence of France for more than a decade. That influence was at first limited and indirect, then pervasive and overpowering. Yet it was during this period of alien preponderance that Germany for the first time felt the stirrings of liberalism and nationalism. The regions that had become part of the French Empire experienced at first hand the advantages of efficient centralized government in which equality before the law and freedom of opportunity were accepted principles. Those states that retained a pseudo-



Germany in 1807 after reconstruction by Napoleon.

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London

independence as satellites of Napoleon, moreover, sought to imitate the example of their master, partly in order to gain his favour, partly in order to emulate his success. One government after another began to remove religious disabilities, relax economic restrictions, eliminate servile obligations, and centralize administrative functions. Above all, constitutional rule and popular representation ceased to seem Utopian to men of property and education who had witnessed the stirring events of the years since 1789. The French hegemony also led to the birth of nationalism in central Europe. For one thing, the achievement of political unity became a distinct possibility, once the territorial fragmentation of the Holy Roman Empire had come to an end. The presence of foreign invaders, furthermore—arrogant, overbearing, and avaricious—aroused among the Germans a sense of nationality that they had never felt in the tranquil days of the old order. Finally, an example of what great deeds the love of fatherland could inspire lay before all who admired or envied the triumphs of Napoleon. The ideal of cosmopolitan individualism that had been generally accepted in the 18th century began to give way before a growing consciousness of national identity. Yet the fact that the concepts of constitutional freedom and national unity were not indigenous but arose in response to foreign domination had an important effect on the form they assumed in central Europe.

Every German state felt the influence of the new principles of government and economy that the period of French hegemony had introduced, but nowhere was that influence more profound or fruitful than in Prussia. For only in the hour of deepest humiliation did the Hohenzollern kingdom finally make an effort to adapt its structure to the changing political and social conditions that it had stubbornly ignored during the years of greatness. Between 1806 and 1813 the statesmen in Berlin initiated a revolution from above in order to transform a rigid despotism into a popular monarchy supported by the loyalty of a free citizenry. Out of the disasters of Jena and Tilsit emerged a group of gifted reformers who sought to

The Prussian reformers: Stein and Hardenberg

prepare the way for the regeneration of their country. The leading figures in this movement for civic reconstruction were the civil servants Karl vom Stein and Karl August von Hardenberg and the military commanders Gerhard von Scharnhorst and August Neithardt von Gneisenau. Among their most important achievements was the abolition of serfdom, a measure designed to create citizens out of human beasts of burden. Yet, while it gave the peasant personal freedom, the government failed to provide him with economic independence. Most of the land remained in the hands of the aristocracy, which therefore continued to dominate the countryside politically as well as socially. More successful was the law establishing municipal self-government. Thereafter, the cities of the kingdom were to be administered by officials chosen not by the central bureaucracy but by the propertied inhabitants of the cities themselves. The autonomy of the cities, it was hoped, would help train a politically conscious and active middle class. The most effective reforms, however, were those introduced in the armed forces. After the officers who had shown themselves incompetent during the war were dismissed or retired, the high command carried out a thorough reorganization of the military system. Discipline became more humane, promotion was made a reward for merit, the method of recruitment was improved, and the training in tactics was modernized. Most important, the army's leaders sought to instill in the soldiers a new spirit rooted in inner conviction rather than unquestioning obedience. Defeat had changed Prussia from a garrison state into a centre of political and intellectual ferment.

Liberalism and nationalism became increasingly vehement in Germany as the burden of French domination grew progressively heavier. The financial sacrifices occasioned by the subordination to Napoleon reinforced the personal resentments aroused by his ruthless statecraft. Before long a network of secret organizations had sprung up in central Europe seeking the expulsion of the foreign invaders. Yet it would be a mistake to think that all Germans regarded the hegemony of France as an unmitigated evil. There were in fact wide differences of opinion among them. The rulers of the secondary states and their supporters in the army and the bureaucracy saw in Napoleon the instrumentality of their new importance. Many reformers in the south—the Bavarian statesman Maximilian von Montgelas, for example—believed that only French influence had made possible the modernization of government in Germany. Some publicists continued to argue, moreover, that the political disunity of central Europe was a natural result of its historic experience and reflected its essential character. To be sure, those who opposed alien domination outnumbered those who accepted it. But even among the former there was no agreement regarding the future political structure of the nation. Many of them dreamed of a liberal and united fatherland that would take its place among the great powers of Europe. Others were willing to settle for a loose association of governments, similar to the Confederation of the Rhine, which could safeguard the interests of the secondary states against Prussia and Austria. Still others hoped for a complete restoration of the old order in which they had grown up and to which they longed to return. And then there were the broad masses of the population of central Europe, exploited, illiterate, and uninformed. They remained by and large indifferent to the crosscurrents of political thought, seeking nothing more than an improvement in their standard of living and the preservation of their way of life. Germany was beginning to move toward new norms of civic and social value, but the transformation of political attitudes was gradual and intermittent.

That the growth of the ideals of unity and freedom was slow became apparent during the first serious effort to throw off the yoke of foreign domination in central Europe. The Austrian government concluded in 1809 that the reverses that Napoleon had been encountering in Spain presaged a general uprising against the French hegemony on the Continent. The result was an ill-fated attempt at a war of liberation, in which the Habsburg troops challenged Napoleon for the fourth time, only to go down in defeat once again. Appeals from Vienna to the people of Germany found little response except in Tirol and among a few nationalist hotspurs in the north. The princes refused to risk French wrath until they could be sure of ultimate victory, while their subjects refused to rise against French oppression without princely approval. The result was that the war in central Europe, unlike the one in the Iberian Peninsula, was waged primarily by regular forces rather than by guerrilla bands. Archduke Charles gained important successes for the Austrian army at Aspern and Essling (May 21–22, 1809), an indication that the strategic mastery of the French was drawing to a close. But at Wagram (July 5–6) Napoleon was able to work the last of his military miracles. Vienna had to sue for peace once more, the Treaty of Schönbrunn (October 14) ceding Salzburg to Bavaria, West Galicia to the grand duchy of Warsaw, and the Adriatic coastland to France. The defeat finally persuaded the emperor, who had exchanged the title Francis II of the Holy Roman Empire for Francis I of the Austrian Empire, that resistance would be as futile in the future as it had been in the past. He therefore adopted a policy of collaboration with France signalized by the marriage of his daughter Marie-Louise to Napoleon. Germany continued to languish in the grip of foreign domination.

The defeat of Austria

The wars of liberation. A new struggle for liberation opened three years later with the defeat of Napoleon's *grande armée* in Russia. As the tsarist armies began to cross their western frontiers in December 1812, the crucial question became what reception they would find among the rulers and the inhabitants of central Europe. The first state to cut its ties to Paris was Prussia. It was not the king, however, but one of his generals, Ludwig Yorck von Wartenburg, who decided on his own initiative to cooperate with the Russians. Only hesitatingly and fearfully did Frederick William III then agree in February 1813 to a war against France, although public opinion in his kingdom greeted the outbreak of the conflict with enthusiasm. The other rulers of central Europe refused initially to follow the Prussian example. The members of the Confederation of the Rhine were still convinced of Napoleon's invincibility, while Austria preferred to see the combatants exhaust each other to the point at which it could play the role of mediator and arbiter. The foreign minister in Vienna, Clemens Lothar von Metternich, was afraid that the hegemony of France in central Europe might be replaced by that of Russia. He tried, therefore, to pursue a strategy of armed neutrality, hoping that he could persuade the opposing sides to accept a compromise by which an equilibrium would be maintained between Alexander I and Napoleon. This plan failed because of the obstinacy of the latter, who feared that concessions in foreign affairs would weaken his control over internal politics in France. The upshot was that in August 1813 Austria entered the conflict on the side of Russia and Prussia, and the balance of military power shifted in favour of the anti-French coalition. The faith of the secondary states in Napoleon's star began to weaken, and Bavaria became the first member to secede from the Confederation of the Rhine (October 8). One great allied victory would now suffice to bring all of Germany into the struggle against France.

The anti-French alliance of Russia, Prussia, and Austria

That victory came on Oct. 16–19, 1813, at the Battle of Leipzig. After four days of bitter fighting, the French army was forced to retreat, and its domination of central Europe was finally at an end. Before the year was out, Napoleon had withdrawn across the Rhine. Of all his conquests in Germany, only the left bank was still under the effective control of Paris. The Confederation of the Rhine promptly collapsed, as its members rushed to go over to the winning side before it was too late. The Rhineland was also reconquered early in 1814, after the allies had launched their invasion of France. In the course of the spring the capture of Paris, the restoration of the Bourbons, and the conclusion of peace in the first Treaty of Paris (May 30) ended the war of liberation except for the episode of the Hundred Days, when Napoleon briefly returned to power and was ultimately and finally beaten at Waterloo. The western frontier of central Europe was to remain essentially the same as at the time of the initial outbreak of hostilities more than 20 years before. New

state boundaries within Germany would still have to be determined, to be sure, and the problem of a new political organization of the nation awaited the victorious statesmen, but the period of foreign hegemony was over at last. The rulers of central Europe, relying partly on the forces of innovation, partly on those of tradition, had succeeded in freeing themselves from alien domination. Now they had to decide what use they would make of their freedom. Would they create a new polity of unity and liberty, which many reformers demanded, or would they reestablish the old order of absolutism and particularism, which the conservatives advocated? As the statesmen began to gather in Vienna in the fall of 1814 to restore peace to a continent ravaged by two decades of war, they pondered the problem of devising an enduring form of government for Germany.

Results of the Congress of Vienna. The men who, in the nine months from September 1814 to June 1815, redrew the map of Europe were diplomats of the old school. Francis I and Prince von Metternich of Austria, Frederick William III and Prince von Hardenberg of Prussia, Alexander I of Russia, Viscount Castlereagh of England, Talleyrand of France, and the representatives of the secondary states were all intellectual heirs of the 18th century. They feared the principles of the French Revolution, they scorned the theories of democratic government, and they opposed the doctrines of national self-determination. But they recognized that the boundaries and governments of 1789 could not be restored without modification or compromise. There had been too many changes in attitudes and loyalties that the rigid dogmas of legitimism were powerless to undo. The task before the peacemakers was thus the establishment of a sound balance between necessary reform and valid tradition capable of preserving the tranquillity that Europe desperately needed. The decisions regarding Germany reached during the deliberations in Vienna followed a middle course between innovation and reaction, avoiding extreme fragmentation as well as rigid centralization. The Confederation of the Rhine was not maintained, but neither was the Holy Roman Empire restored. Although the reforms introduced during the period of foreign domination were partly revoked, the practices of enlightened despotism were not entirely reestablished. Despite the complaints of unbending legitimists and the dire predictions of disappointed reformers, the peacemakers succeeded in creating a new political order in central Europe that endured for half a century. The long years of war and unrest that had convulsed Europe during the era of the French Revolution and Napoleon were followed by even longer years of stability and tranquillity.

The Germany that emerged in 1815 from the Congress of Vienna included 39 states ranging in size from the two great powers, Austria and Prussia, through the minor kingdoms of Bavaria, Württemberg, Saxony, and Hannover; through smaller duchies such as Baden, Nassau, Oldenburg, and Hesse-Darmstadt; through tiny principalities such as Schaumburg-Lippe, Schwarzburg-Sondershausen, and Reuss-Schleiz-Gera; to the free cities of Hamburg, Bremen, Lübeck, and Frankfurt am Main. The new boundaries in central Europe bore little resemblance to the bewildering territorial mosaic that had been maintained under the Holy Roman Empire, but there were still many fragments, subdivisions, enclaves, and exclaves, too many for the taste of ardent nationalists. Yet the overall pattern of state frontiers represented a significant improvement over the chaotic patchwork of sovereignties and jurisdictions that had characterized the old order. The peacemakers not only created more integrated and viable political entities but also altered the role that these entities were to play in the affairs of the nation. Without design or even awareness on the part of Frederick William III, his kingdom of Prussia assumed a pivotal position in Germany. The victorious powers, on guard against a revival of French aggression, decided to make Berlin the defender of the western boundary of central Europe. The Rhineland and Westphalia, a region destined to develop into the greatest industrial centre on the Continent, became Hohenzollern provinces. More than that, the king agreed at the urging of Alexander I to cede the bulk of his Polish possessions to Russia in return for a substantial part of Saxony. Prussia, which at the end of the 18th century had been in the process of becoming a binational state, was thrust back into Germany and given a strategic position on both frontiers of the nation. The centre of gravity of Austria, on the other hand, shifted eastward. Francis I had decided to abandon the historic role of his state as protector of the Holy Roman Empire against the Bourbons for the sake of greater geographic compactness and military defensibility. The possessions in southern and western Germany were surrendered along with the Austrian Netherlands in return for Venetian territory on the Adriatic. The Habsburg empire thus became less German in composition and outlook as its focus shifted in the direction of Italy and the Balkans. The consequences of this territorial rearrangement were to be far-reaching.

Germanic Prussia and the polyglot Austrian Empire



Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London

The German Confederation, 1815.

THE AGE OF METTERNICH AND THE ERA OF UNIFICATION: 1815–71

Reform and reaction. In place of the Holy Roman Empire the peacemakers of the Congress of Vienna had established a new organization of states of central Europe, the German Confederation. This was a loose political association in which most of the rights of sovereignty remained in the hands of the member governments. There was no central executive or judiciary, only a federal Diet meeting in Frankfurt am Main to consider common legislation. The delegates who participated in its deliberations were representatives appointed by and responsible to the rulers whom they served. The confederation was in theory empowered to adopt measures strengthening the political and economic bonds of the nation. In fact it remained a stronghold of particularism, unwilling to sacrifice local autonomy in order to establish centralized authority. It was designed essentially to defend the interests of the secondary states and the Habsburgs. The former, jealously guarding the independence and importance they had gained during the period of French hegemony, were opposed to any reform that might limit their sovereignty. The latter believed that only a decentralized form of political union in Germany would give them enough freedom of action to pursue their non-German objectives. The confederation was thus from the outset an ally of localism and traditionalism. To the nationalists, whose hopes had

risen so high during the war of liberation, it seemed to be an instrument of blind reaction. Yet the truth is that the confederal system established in 1815 accurately reflected the slow development of civic consciousness and economic integration in central Europe. The militant reformers who demanded the centralization of government were a vocal but small minority. The lower classes accepted the territorial and constitutional decisions of the Congress of Vienna without a murmur of protest. The weakness of the peace settlement was not its failure to embody present realities but its inability to adjust to future changes. What had been a reasonable adaptation to the political needs of an agrarian and rural society became a hopeless anachronism 50 years later in the age of factories and railroads. This was the fatal flaw in the German Confederation.

Major flaw of the German Confederation

Yet the reform movement that had begun under the impact of the French hegemony did not end with the downfall of Napoleon. It continued to exert influence over affairs of state for another few years, before the forces of authoritarianism and particularism crushed it. That influence was strongest in southern Germany, where the political example of western Europe had made the deepest impression. There, many civil servants, court officials, army officers, and even aristocratic landowners came to believe that the future of the state depended on its readiness to reform civic institutions in accordance with liberal theories. In the years following Waterloo, one government in the south after another promulgated a constitution: Bavaria and Baden in 1818, Württemberg in 1819, and Hesse-Darmstadt in 1820. These constitutions established representative assemblies, elected by the propertied citizens, whose assent was required for the enactment of legislation. Their purpose was not only to win for the crown the support of the educated classes of society but also to engender a sense of unity in a heterogeneous population that still had diverse allegiances and traditions. To the north there were also persistent echoes of the reform movement.

The followers of Karl vom Stein were still influential in the councils of state, and Frederick William III of Prussia at first seriously considered ways of fulfilling the promise he had made in 1815 to establish constitutional government. The agitation for political reorganization was loudest, however, among university students, who formed patriotic groups known as *Burschenschaften*. They demanded the abandonment of the confederal system, the establishment of greater unity, and the achievement of national power. Gathering in 1817 at the Wartburg, a castle near Eisenach, they listened to veiled denunciations of the existing order and consigned to flames various symbols of traditional authority. The rulers of Germany began to stir uneasily at this bold display of defiance of legitimate government.

The chief strategist of the forces hostile to reform was Metternich. Not only did he reject the teachings of liberalism and nationalism in principle, but also, as the leading statesman of the Habsburg empire, he recognized that the establishment of centralized authority in Germany would seriously impede the policies his government was pursuing in Hungary, Italy, and the Balkans. When on March 23, 1819, an unbalanced student, Karl Ludwig Sand, assassinated the conservative playwright and publicist August von Kotzebue, Vienna persuaded the princes of the German Confederation that they were facing a dangerous attempt to overthrow the established order in central Europe. The result was a series of repressive measures called the Carlsbad Decrees, which the federal Diet adopted on Sept. 20, 1819. General censorship was introduced, and the *Burschenschaften* were outlawed. This first major success of the conservative counteroffensive had an important effect on the struggle within the state governments between the advocates and the opponents of reform. In Prussia the liberal members of the ministry were forced to resign, and the plan to promulgate a constitution for the kingdom was rejected. This shift to the right by Berlin encouraged authoritarian tendencies among the secondary states of the north, which soon abandoned their own constitutional projects. By the end of 1820 the reform movement, which had begun some 15 years before, came to a complete halt.

Repression and the Carlsbad Decrees

It had succeeded in altering the political and economic structure of society, but it had been unable to establish a tradition of liberal government and national loyalty in central Europe. The forces of particularism and legitimism, deriving their chief support from the landowning nobility and the conservative peasantry, remained strong. The foundation of bourgeois civic consciousness and material prosperity on which England and France had built their representative institutions was still lacking beyond the Rhine. The ideal of free government was introduced in Germany not as the fruit of an industrial or a political revolution but as the imitation of a foreign example and the reaction against a foreign oppression.

The established order was once again threatened briefly in the wake of the July days of 1830 in France. The news that there had been a successful insurrection against the Bourbons in Paris had an electrifying effect throughout the Continent. In central Europe there were sympathetic uprisings in some of the secondary states of the north. The rulers of Brunswick, Saxony, Hannover, and Hesse-Kassel, seeking to forestall more extreme demands, agreed to promulgate liberal constitutions. A mass meeting of southern radicals at Hambach Castle in the Palatinate (May 1832), moreover, expressed its approval of national unification, republican government, and popular sovereignty. A group of militant students even launched a foolhardy attempt to seize the city of Frankfurt am Main, dissolve the federal Diet, and proclaim a German republic. The effect of such harebrained schemes was predictable. As the princes of the German Confederation gradually recovered from their initial fear of the revolutionary movement, they began to oppose with increasing vigour plans to alter the existing system of government. Again, Metternich took the lead in the effort to crush liberalism and nationalism. Under his direction the federal Diet adopted additional repressive measures reinforcing the position of the crown in state politics, limiting the power of the legislature, restricting the right of assembly, enlarging the authority of the police, and intensifying the censorship. Within a few years the opposition had been subdued, and the German Confederation could continue to vegetate in its cozy provincialism. Not until the middle years of the century did a new and more violent outburst of political disaffection shake the foundations of the system of authority that the Congress of Vienna had erected.

Effects of the July Revolution of 1830

Evolution of parties and ideologies. Although the critics of the established order could be defeated, they could not be silenced. The struggle between the supporters and the adversaries of the existing form of government led to the emergence of a rudimentary party system in the German Confederation. In the legislative assemblies of the secondary states the proponents of reform began to meet, plan, organize, and propagandize. The defenders of legitimism were thereby forced in turn to concert their strategy and to publicize their program. Even in Prussia and Austria, where there were as yet no constitutions or parliaments, political criticism could be expressed obliquely through clubs, meetings, newspapers, pamphlets, and petitions. The result was the gradual development of amorphous civic associations held together by common convictions regarding the nature of state and society. These primitive groupings were only the raw material out of which disciplined political parties were slowly fashioned in the course of the century. They still lacked a clear sense of purpose and the systematic propagation of belief characteristic of a fully mature system of parliamentary politics. Yet they became the instrumentalities by which disaffected groups in the community could express their opposition to the established order. They reflected the fact that the civic attitudes of the period of the Restoration were no longer those of the age of enlightened despotism. There were men in central Europe now who refused to submit without question to princely authority, seeking freedom only in the inner recesses of the soul. The change in the form of economy and the structure of society produced by the beginnings of industrialization led to an alternation in the system and organization of politics.

The most important opponents of legitimism and particularism were the liberals or moderates. Deriving their sup-

The liberals or moderates

port primarily from industrialists, merchants, financiers, mineowners, railroad promoters, civil servants, and university professors, they represented the opposition of the well-to-do bourgeoisie to a form of government in which an aristocracy of birth rather than of talent predominated. Their political principles favoured a monarchical system of authority, but the crown was to share its powers with a parliament elected by the propertied classes. Influence in public affairs should be accessible to all citizens who had demonstrated through the acquisition of wealth and education that they were capable of exercising the franchise intelligently. While the liberals resented the inherited privileges of the nobility, they also feared the proletariat. The man who lived in poverty and ignorance, they reasoned, was ripe for demagoguery and insurrection. The path of civic wisdom was therefore the *juste-milieu* between royal absolutism and mob rule, which had been established in England by the Reform Bill of 1832 and in France by the regime of Louis-Philippe. In economics, the teachings of liberalism advocated a policy of unrestrained competition by which wealth would become the reward of business acumen rather than the perquisite of corporative privilege. Guild monopolies, government regulations, and entailed estates were to be abolished as violations of that freedom of enterprise that alone could ensure the well-being of all society. The liberals favoured the transformation of the German Confederation into a national monarchy in which the states' rights would be curtailed but not destroyed by a central government and a federal parliament.

The democrats or radicals

Farther to the left stood the democrats or radicals, whose following was made up largely of small businessmen, petty shopkeepers, skilled workers, independent farmers, publicists, journalists, lawyers, and physicians. They looked with scorn on the golden mean between autocracy and anarchy that the liberals were seeking. They preferred an egalitarian form of authority in which not parliamentary plutocracy but popular sovereignty would be the underlying principle of government. Their supporters drew inspiration from the French rather than the English or the American Revolution. Their ideal of petty bourgeois democracy was the Jacobin republic of 1793 as an instrument to shape the energies and aspirations of the people into a disciplined force for political and social reform. The spokesmen of this ideal could not openly demand the overthrow of monarchical institutions without risking imprisonment. Yet, while they were forced to accept the crown as a political institution, they sought to transfer its power to a parliament elected by equal manhood suffrage. The masses would thereby become the ultimate arbiter of politics. The democrats were also willing to accept government regulation of business activity as a means of improving the economic position of the lower classes, although their belief in the sanctity of private property was as firm as that of the liberals. In their advocacy of national unification, however, they were less solicitous about royal prerogatives and state rights. While not as numerous as the moderates, the radicals remained an important source of opposition to the established order.

The conservatives or legitimists

The growth of criticism directed against the political system of the Restoration forced its supporters to define their ideological position with greater precision. The old theories of monarchy by divine right or despotic benevolence offered little protection against the assaults of liberalism and democracy. Legitimism, whose defenders came mostly from the landed nobility, the court aristocracy, the officer corps, the upper bureaucracy, and the established church, began therefore to advance new arguments based on conservative assumptions about the nature of man and society. The relationship between the individual and government, so the reasoning went, cannot be determined by paper constitutions founded on a doctrinaire individualism. Human actions are not motivated solely by rational considerations but by habit, feeling, instinct, and tradition as well. The impractical theories of visionary reformers fail to take into account the historic forces of organic development by which the past and the present shape the future. To assert that all men are equal is to ignore differences in rights and duties expressing differences in birth, class, background, education, and tradition. The dogmas of constitutional authority and parliamentary government are merely a facade behind which a self-seeking bourgeoisie seeks to disguise its lust for power. An enduring form of government can be built only on the traditional institutions of society: the throne, the church, the nobility, and the army. Only a system of authority legitimated by law and history can protect the worker against exploitation, the believer against godlessness, and the citizen against revolution. According to these tenets, the political institutions of the German Confederation were valid, because they represented fundamental convictions deeply embedded in the spirit of the nation.

Economic changes and the Zollverein. The struggle of parties and ideologies during the Restoration reflected far-reaching changes in the structure of the economy and the community. The most significant of these changes was the rise of large-scale industry in central Europe. Techniques of mechanization, introduced in textile mills and coal mines, spread to other branches of manufacture and exerted pressures that influenced the entire economic life of the nation. The transportation network improved with the construction of railroads, steamships, bigger highways, and better canals. Banking institutions and private investors began to transfer their funds from government bonds and commercial ventures to manufacturing enterprises. Millowners, ironmasters, railroaders, financiers, and stockbrokers gradually formed a new middle class whose wealth was derived primarily from industrial activity and whose growing economic importance encouraged among its members a demand for greater political influence. Skilled handicraftsmen, constituting the bulk of the urban working class, could not compete successfully with the factories. The conflict between industrial and pre-industrial forms of output was aggravated, moreover, by important demographic changes in the period following the Congress of Vienna. Population began to shift from country to city, although a majority of the inhabitants of the German Confederation continued to live in rural communities.

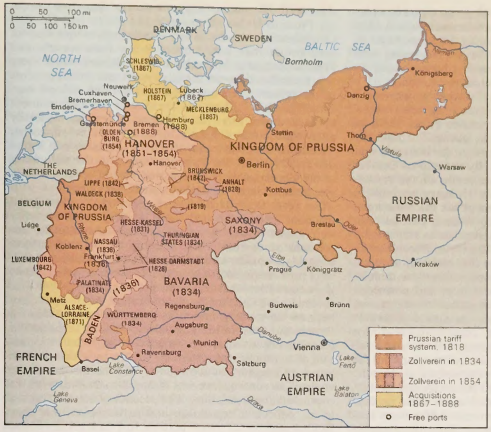
The growth of large-scale industry

Agriculture also went through as difficult a period of reorganization and rationalization as industry. The end of serfdom had led in the regions east of the Elbe to the expansion of large estates belonging to aristocratic landowners but cultivated by a propertyless rural proletariat. Peasant emancipation in Prussia allowed the Junkers to enlarge their lands by absorbing the holdings of small farmers. The result was the continuing economic, social, and political domination of the village by the nobility in the eastern provinces of the Hohenzollern kingdom. The squirearchy entrenched in Pomerania, Brandenburg, Silesia, and East Prussia controlled agriculture, commanded the army, directed the bureaucracy, and influenced the court. It constituted a powerful force for conservatism and particularism.

The reorganization of agriculture

West of the Elbe the basic problem was not landlessness but overpopulation. The aristocracy along the Rhine and the Danube was often willing to give the peasantry possession of the soil in return for a substantial payment. The farmer was thereby saddled with heavy financial obligations. Many rustics tried to escape poverty by emigrating to the New World; those who remained faced swift demographic expansion and often had to subdivide small holdings until they yielded no profit. Civil discontent mounted among impoverished villagers who lacked employment in industry.

The system of authority in the German Confederation was thus being undermined by the struggle of artisans against industrial mechanization, by the disaffection of peasants hungry for land, and above all by the criticism of businessmen groaning under the shackles of particularism. Industrialists and financiers had to overcome the barriers created by a variety of monetary systems, commercial regulations, excise taxes, and state boundaries. It is little wonder that the bourgeoisie of central Europe turned increasingly to the teachings of liberalism and nationalism. Yet the established order did make a major attempt to meet the needs of the business community. Before long, several of the more important secondary governments concluded agreements with Berlin by which a sizable free



The growth of the German Zollverein.

The Zollverein

trade area was established in the heart of central Europe.

In 1834 the Zollverein, or Customs Union, including most of the states of the German Confederation, came into existence. Only Austria and the northwest coastland remained aloof. The industrialists of the Habsburg empire, who wanted their products protected against outside competition, felt that the tariff rates of the new association were too low for their needs, whereas the merchants and bankers of the coastal region, who depended on imports, thought that they were too high. Yet for some 25 million Germans the Zollverein meant in effect the achievement of commercial unification without the aid of political unification. The Prussian government, moreover, acquired a powerful new weapon in the struggle against Austria for the dominant position in central Europe. Still, although the Zollverein helped meet the most pressing demands of the middle class for economic consolidation, it could not surmount all the material disadvantages of a particularistic form of government. People of means and education continued to grumble about the confederal system under which they lived, while the masses became increasingly restless under the pressure of social dislocation.

The revolutions of 1848–49. The hard times that swept over the Continent in the late 1840s transformed widespread popular discontent in the German Confederation into a full-blown revolution. After the middle of the decade a severe business depression halted industrial expansion and aggravated urban unemployment. At the same time, serious crop failures led to a major famine in the entire area from the Irish Sea to Russian Poland. In central Europe the hungry '40s drove the lower classes, which had long been suffering from the economic effects of industrial and agricultural rationalization, to the point of open rebellion. There were sporadic hunger riots and violent disturbances in several of the states, but the signal for a concerted uprising did not come until early in 1848 with the exciting news that the regime of the bourgeois king Louis-Philippe had been overthrown by an insurrection in Paris (February 22–24). The result was a series of sympathetic revolutions against the governments of the German Confederation, most of them mild but a few, as in the case of the fighting in Berlin, bitter and bloody. When on March 13 Metternich, the proud symbol of the established order, was forced to resign his position in the Austrian Cabinet, the princes hastened to make peace with the opposition in order to forestall republican and socialist experiments like those in France. Prominent liberals were appointed to the state ministries, and civic reforms were introduced safeguarding the rights of the citizens and the powers of the legislature. But even more important was the attempt to achieve political unification through a national assembly representing all of Germany. Elections were held soon after the spring uprising had subsided, and on May 18 the Frankfurt National Assembly met to prepare the constitution for a free and united fatherland. Its convocation represented the realization of the hopes that nationalists had cherished for more than a generation. Within the space of a few weeks, those who had fought against the particularistic system of the Restoration for so long suddenly found themselves in power with a popular mandate to rebuild the foundations of political and social life in central Europe. It was an intoxicating moment.

The Frankfurt National Assembly

The forces that had defeated the establishment order, however, soon discovered that they were in disagreement regarding the use to be made of their common victory. There were, first of all, sharp differences between the liberals and the democrats. While the former had comfortable majorities in most of the state legislatures as well as in the Frankfurt parliament, the latter continued to plead, agitate, and conspire for a more radical course of action. There were also bitter disputes over the form that national unification should assume. The Grossdeutsch (Great German) movement maintained that only Austria, the state whose rulers had worn the crown of the Holy Roman Empire for 400 years, could successfully guide the united fatherland. The Kleindeutsch (Little German) party, on the other hand, argued that the Habsburgs had too many Slavic, Magyar, and Italian interests to work single-mindedly for the greatness of Germany. The natural leader of the nation was Prussia, whose political vigour and geographic position would provide efficient government and military security for central Europe. Finally, there was a basic conflict between the proletariat, which wanted protection against mechanized production and rural impoverishment, and the bourgeoisie, which sought to use the political power it had gained in order to promote an industrial capitalism based on freedom of enterprise. Once the spring uprising was over, the parties and classes that had participated in it began to quarrel about the nature of the new order that was to take the place of the old. Popular support for the revolution, which had made the defeat of legitimism during the March days possible, began to dwindle, discouraged by the realization that the liberals would do no more to solve the problems of the masses

than the conservatives had done. While the Frankfurt parliament was debating the constitution under which Germany would be governed, its following diminished and its authority declined. The forces of the right, recovering from the demoralization into which they had been plunged by the revolution, began to plan a counterrevolution.

Suppression of the revolutions

Their first major victory came in Austria, where the young emperor Francis Joseph found an able successor to Metternich in his prime minister Prince Felix zu Schwarzenberg. In the course of the summer of 1848 the Habsburg armies crushed the uprising in Bohemia and checked the insurrection in Italy. By the end of October they subjugated Vienna itself, the centre of the revolutionary movement, and now only Hungary was still in arms against the imperial government. At the same time, in Prussia the irresolute Frederick William IV had been gradually persuaded by the conservatives to embark on a course of piecemeal reaction. Early in December he dissolved the constituent assembly that had been meeting in Berlin, promulgated by his own authority a middle-of-the-road constitution for the kingdom, and proceeded little by little to reassert the prerogatives of the crown. Among the secondary states there was also a noticeable shift to the right, as particularist princes and legitimist aristocrats began to regain their courage. By the time the Frankfurt parliament completed its deliberations in the spring of 1849, the revolution was everywhere at ebb tide. The constitution that the National Assembly had drafted provided for the creation of a federal union at the head of which was to stand a hereditary emperor with powers limited by a popularly elected legislature. Since the Austrian government had already indicated that it would oppose the establishment of a centralized system in Germany, the imperial crown was offered to the king of Prussia. Frederick William IV hesitated, brooded, and agonized, but in the end he refused the dignity whose authority was in his opinion too restricted. This rejection of political consolidation under a liberal constitution destroyed the last chance of the revolutionary movement for success. The moderates, admitting failure, went home to mourn the defeat of their hopes and labours. The radicals, on the other hand, sought to attain their objectives by inciting a new wave of insurrections. Their appeals for a mass uprising, however, were answered mostly by visionary intellectuals, enthusiastic students, radical politicians, and professional revolutionaries. The lower classes remained by and large indifferent. There was sporadic violence, especially in the southwest, but troops loyal to princely authority had little difficulty in defeating petty bourgeois democracy. By the summer of 1849 the revolution, which had begun a year earlier amid such extravagant expectations, was completely crushed.

The 1850s: years of political reaction and economic growth. The attempt to achieve national unification through liberal reform was followed by an attempt to achieve it through conservative statesmanship. Frederick William IV had refused to accept an imperial crown vitiated by parliamentary government, but he was willing to become the head of a national federation in which the royal prerogative remained unimpaired. While the Austrian armies were still engaged in the campaign against the revolution in Hungary, Berlin began to exert diplomatic pressure on the secondary governments to join in the formation of a new federal league known as the Prussian Union. If Frederick William IV had acted with enough determination, he might have been able to reach his goal before Francis Joseph could intervene effectively in the affairs of Germany. But he allowed his opportunity to slip away. Though he succeeded through threats and promises in persuading most of the princes to accept his proposals, no irrevocable commitments had been made by the time the Hungarians were defeated in August 1849. Vienna could now proceed to woo the secondary governments, which had in most cases submitted to Prussia only out of weakness and fear. Basically they remained opposed to sacrificing their sovereignty in order to exalt the Hohenzollern dynasty. When Schwarzenberg suggested the re-establishment of the old federal Diet, he won the support of many rulers who had agreed to follow Berlin against their will. The nation was now divided into two camps, the Prussian Union on one side and the revived German Confederation on the other. It was only a question of time before they would clash. When both Austria and Prussia decided to intervene in Hesse-Kassel, where there was a conflict between the supporters and the opponents of the prince, Germany stood on the brink of civil war. But Frederick William IV decided at the last moment to back down. His fear overcame his pride, especially after Nicholas I of Russia indicated that he supported Vienna in the controversy. By the Punctation of Olmütz of Nov. 29, 1850, the Prussians agreed to the restoration of the German Confederation, and the old order was fully reestablished in all its weakness and inadequacy.

Revival of the German Confederation

The years that followed were a period of unmitigated reaction. Those who had dared to defy royal authority were forced to pay the penalty of harassment, exile, imprisonment, or even death. Many of the political concessions made earlier, under the pressure of popular turmoil, were now restricted or abrogated. In Austria, for example, the constitution that had been promulgated in 1849 was revoked, and legitimacy, centralization, and clericalism became the guiding principles of government. While in Prussia the constitution granted by the king remained in force, its effectiveness was reduced through the introduction of a complicated system of election by which the ballots were weighted in accordance with the income of the voters. The consequence was that well-to-do conservatives controlled the legislature. The secondary states returned to the policies of legitimism and particularism that they had pursued before the revolution. In Frankfurt am Main, where the federal Diet now resumed its sessions, diplomats continued to guard the prerogatives of princely authority and state sovereignty. The restoration of the confederal system also served the interests of the Habsburgs, who stood at the pinnacle of their prestige as the saviours of the established order. In Berlin, on the other hand, the prevailing mood was one of confusion and discouragement. The king, increasingly gloomy and withdrawn, came under the influence of ultraconservative advisers who preached legitimism in politics and orthodoxy in religion. The government, smarting under the humiliation suffered at the hands of Austria, was as timid in foreign as it was oppressive in domestic affairs. The people, tired of insurrection and cowed by repression, were politically apathetic. The German Confederation as a whole, rigid and unyielding, remained during these last years of its existence blind to the need for reform that the revolution had made clear.

Industrial capitalism

Yet the 1850s, so barren in politics, were of the highest importance in economics, for it was during this period that the great breakthrough of industrial capitalism occurred in central Europe. The national energies, frustrated in the effort to achieve civic reform, turned to the attainment of material progress. The victory of the reaction was followed by an economic expansion as the business community began to recover from its fear of mob violence and social upheaval. The influx of gold from America and Australia, moreover, generated an inflationary tendency, which in turn encouraged a speculative boom. Not only did the value of industrial production and foreign trade in the Zollverein more than double in the course of the decade, but also new investment banks based on the joint-stock principle were founded to provide risk capital for factories and railroads. The bubble burst in 1857 in a financial crash that affected the entire Continent. For many investors the price of overoptimism and speculation turned out to be misfortune and bankruptcy. Yet Germany had now crossed the dividing line between a preindustrial and an industrial form of economy. Although the rural population still outnumbered the urban, the tendency toward industrialization and urbanization had become irreversible. And this in turn had a profound effect on the direction of politics. For, as wealth continued to shift from farming to manufacturing, from the country to the city, and from the aristocracy to the bourgeoisie, the pressure for a redistribution of political power also gained strength. While the reactionaries were solemnly proclaiming the sanctity of traditional institutions, economic change was undermining the foundation of those institutions. By the

end of the decade a new struggle between the forces of liberalism and conservatism was in the making.

The 1860s and the triumphs of Bismarck. The revival of the movement for liberal reform and national unification at the end of the 1850s came to be known as the "new era." Its coming was heralded by scattered but distinct indications that the days of the reaction were numbered. In 1859 the defeat of Austria in the war against France and Piedmont had a profound effect on central Europe. For one thing, the maintenance of the authoritarian regime in Vienna depended on respect for its military strength. Now that the Habsburg armies had shown themselves to be vulnerable, popular unrest in the empire began to increase. Since autocracy was no longer an effective principle of government, Francis Joseph decided to experiment with a parliamentary form of authority. On Oct. 20, 1860, he promulgated a constitution (the October Diploma) for his domains, setting up a bicameral legislature with an electoral system favouring the bourgeoisie, and Austria ceased to be an absolutist state. The achievement of political consolidation in Italy, moreover, aroused hope and envy north of the Alps. If the Italians could overcome the obstacles of conservatism and particularism, why not the Germans? National sentiment in central Europe, dormant since the revolution, suddenly awoke. Patriotic organizations like the Nationalverein (National Union) and the Reformverein (Reform Union) initiated an agitation for a new federal union, the former advocating Prussian, the latter Austrian, leadership. Liberal publicists and politicians began to advance plans for the reconstruction of the German Confederation. Some of the states, detecting a shift in the trend of public opinion, decided to change their course accordingly. Here and there the conservative ministers of the reaction were retired or dismissed, and their place was taken by statesmen with more moderate views.

The most significant portent of a new age in politics, however, appeared in Prussia. In 1857 Frederick William IV, crushed by memories of the mass insurrections and diplomatic defeats that he had been forced to endure, suffered a mental breakdown. A year later his brother became regent, and the government in Berlin immediately began altering the direction of its policy. Prince William, although a man of conservative inclination, had little sympathy with the mystical visions and pious dogmas prevailing at the court during the period of reaction. He dismissed the Cabinet that had served his predecessor, announced a program of cautious reform in Prussian as well as German affairs, and won a popular endorsement of his course in the elections that gave the liberals control of the legislature. After a long period of discouragement, the advocates of civic reconstruction could once again look to the future with hope and expectation.

Yet there was an important difference between the political attitude of the liberals in 1858–59 from that of 1848–49. They came to feel during the new era that their defeat 10 years before had been due to an excess of idealism and exuberance. The fatal mistake of the revolution, they reasoned, had been the assumption that enthusiasm and selflessness could be translated into power and substituted for statesmanship. Now a more calculating policy, a policy of realpolitik, must be adopted. Not theory and rhetoric but negotiation and compromise would lead to the attainment of unity and freedom. The liberals therefore pursued at first a strategy of conciliation, anxious not to frighten the established order into blind resistance against all reform. In Prussia, for example, they waited patiently for the regent to move against the forces of disunity and oppression, confident that if they only gave him enough time he would obtain for them by royal authority what they could not seize through revolutionary violence. Yet it gradually became apparent that their hopes would not be realized. Prince William, who in 1861 became king in his own right, was a moderate conservative but a conservative nevertheless. As the advocates of reform grew increasingly restless, the more militant among them formed the Progressive Party, which sought to hasten the introduction of liberal legislation by exerting pressure on the government. The monarch, afraid that he was being pushed farther to the left than he wanted to go, became more adamant and uncompromising. Sooner or later a conflict between crown and parliament was bound to arise.

The Prussian policy of realpolitik

It came in connection with the question of army reform, although, if that issue had not developed, there would undoubtedly have been another. The king wanted to strengthen the armed forces by increasing the number of line regiments and decreasing the size of the popular militia. The legislature, reluctant to enhance the power of the conservative officer corps, demanded a modification of the plan. The king refused, convinced that the politicians were attempting to gain control of the army, which he considered subject only to royal authority. The legislature responded by withholding approval of the budget to pay for the cost of army reform. A complete deadlock ensued. In the spring of 1862 the liberal ministers who had been appointed after the establishment of the regency were dismissed, and a conservative Cabinet took office. But the new leaders of the government were as unsuccessful as the old in resolving the crisis. William I began to think about abdicating in favour of his son, who was believed to have political views close to those of the parliamentary opposition. He was persuaded, however, to consider first the possibility of naming a new ministry led by Otto von Bismarck, the Prussian ambassador to Paris. There was a momentous interview between the monarch and the envoy, as a result of which the former abandoned all thought of retirement, and the latter became head of a Cabinet pledged to continue the struggle against the legislature. The battle between crown and parliament, which the liberals had been on the point of winning, was now to be waged without regard for constitutional provisions concerning the budget. On Sept. 23, 1862, the nation was startled by the news that a statesman with a reputation for unyielding conservatism had become the prime minister of Prussia. It meant that the established order, having successfully defended its interests against the forces of reform after 1815 and after 1848, was determined to fight to the bitter end against the new challenge to its predominance.

Advent of Otto von Bismarck

The constitutional conflict in the Hohenzollern kingdom continued for another four years. The legislature refused to approve the budget until its wishes concerning military reform had been met. Bismarck's government, after carrying out the controversial reorganization of the army, continued to collect taxes and disburse funds without regard for parliamentary authorization. The liberals condemned the prime minister as a violator of the law, while the prime minister denounced the liberals as blindly doctrinaire. Although the electorate remained on the side of the opposition, the Cabinet declared that it would not be swayed by party politics or parliamentary majorities. The broad masses of the population, it maintained, were still loyal to the crown. And so the struggle went on without prospect of alleviation or resolution. There were even dark prophecies of a violent uprising against a regime that was so indifferent to its constitutional obligations. Yet in fact Bismarck was not blind to the need for a reconciliation between crown and bourgeoisie. Despite his reputation as a fire-eating legitimist, he had a supple mind and recognized that the political principles of Frederick the Great could not solve the problems facing William I. He hoped, therefore, for an eventual reconciliation between the government and the legislature, but a reconciliation in which the prerogative of the monarch and the influence of the nobility would remain undiminished.

What Bismarck sought in essence was an alteration in the form of government that would create a facade of parliamentary institutions disguising the continuation of authoritarian policies. The middle class wanted to end the domination of the traditional forces in society, he calculated, but it also wanted to achieve national unification in central Europe. Here was the key to a solution of the constitutional conflict. Unity could be used to restrict freedom; nationalism could become the means of taming liberalism. Bismarck had concluded that the political integration of Germany was, in the long run, inevitable. If the established order did not effect it, the reformers, democrats, and revolutionaries would. Thus, it was in the interest of conservatism to take the task of centralization in hand, bring it to a successful conclusion, and create a

new system of authority compatible with the preservation of royal and aristocratic preponderance. Such a policy would make possible a compromise between crown and bourgeoisie by which the latter obtained the benefits of economic consolidation while the former retained the advantages of political domination. Through this strategy the prime minister hoped to end the constitutional conflict.

The defeat of Austria. The international situation was favourable to a program of unification in the German Confederation. Since its defeat in the Crimean War (1853–56), Russia had ceased to play a decisive role in the affairs of the Continent. England remained preoccupied with the problems of domestic reform. And Napoleon III was not unwilling to see a civil war east of the Rhine that he might eventually use to enlarge the boundaries of France. Bismarck could thus prepare for a struggle against Austria without the imminent danger of foreign intervention that had faced Frederick William IV. His first great opportunity came in connection with the duchies of Schleswig and Holstein, which were ruled by the king of Denmark but which were politically and ethnically tied to Germany. When the government in Copenhagen sought to make Schleswig an integral part of the Danish state in 1863, nationalist sentiment in central Europe was outraged. William I proposed to Francis Joseph that the two leading powers of the German Confederation should occupy the duchies in order to prevent the violation of an international agreement that had guaranteed their separate status. Afraid to let the Prussians act on their own, the emperor agreed, and in 1864 there was a brief war against Denmark that demonstrated the strength of the reorganized army of the Hohenzollerns. Danish hopes for foreign assistance proved illusory, and by the Peace of Vienna (October 30) the duchies became the joint possession of Prussia and Austria.

The Danish War and Schleswig-Holstein

The easy victory of the allies, however, was only the prelude to a bitter conflict between them. Vienna would have liked to see Schleswig-Holstein become an independent secondary state in the German Confederation, committed to a policy of particularism. Berlin, on the other hand, hoped for the outright annexation of the duchies or at least the indirect control of their government. Even more important than the disposition of the spoils of war, however, was the mounting rivalry between the two great powers for hegemony in central Europe. Bismarck would probably have been willing to collaborate with Austria in the division of Germany into two spheres of influence, a northern under the Hohenzollerns and a southern under the Habsburgs, but Francis Joseph was reluctant to restrict his authority and ambition to the region below the Main River. The result was a steady deterioration of relations between Vienna and Berlin. In 1865 their differences were papered over by the Convention of Gastein, which placed Schleswig under Prussian and Holstein under Austrian administration but which also reaffirmed the joint sovereignty of the two governments over the duchies. Still, this was only a temporary solution, and before long the danger of civil war in the German Confederation began to grow once again.

In the course of the spring of 1866 both sides stepped up their preparations for a military solution to the Austro-Prussian dualism. Bismarck concluded an alliance with Italy by which the latter was to receive Venetia as a reward for participating in a war against the Habsburg empire. He also sought to gain the support of public opinion in the German Confederation by introducing a motion in the federal Diet for the convocation of a national parliament elected by equal manhood suffrage. The Austrians in the meantime secured a promise of French neutrality in the event of hostilities and tried to win the adherence of the secondary states in the impending struggle. The last desperate attempts to preserve peace collapsed in June. Vienna announced that it would submit the question of the duchies to the federal Diet. Berlin, condemning this step as a violation of the Convention of Gastein, ordered its troops in Schleswig to expel the Austrians from Holstein. Francis Joseph in reply called on the other states of the confederation to mobilize their armies against the Prussian threat to domestic tranquillity, and central Europe trembled on the verge of civil war. The only question now was what the position of the secondary states would be. Most of them lined up behind Austria, which they regarded as the defender of their independence against the ambitions of Berlin. Bismarck's attempt to enlist the aid of the national movement by advocating reform of the confederal system thus failed. It alarmed the particularists without propitiating the centralists. Public opinion remained frightened and confused, distrusting one side and fearing the other. The future of the nation was decided not by popular insurrections or parliamentary deliberations but by the force of arms.

The Seven Weeks' War

The Seven Weeks' War between Prussia and Austria (June–August) produced a diplomatic revolution in Europe, destroying the balance of power that had been established 50 years before by the Congress of Vienna. Yet this momentous alteration in the international equilibrium was accomplished so swiftly that foreign diplomats had barely begun to grasp its implications before the struggle for hegemony in Germany ended. The Hohenzollern armies had a brilliant strategist in Helmuth von Moltke and a deadly weapon in the breech-loading needle gun. The Austrian high command, on the other hand, became irresolute and demoralized before a decisive encounter had even taken place. The Prussians succeeded in dividing and defeating the forces of the secondary states, and on July 3 they routed the Habsburg troops as well at the Battle of Königgrätz (Sadowa). The war was thus decided within a few weeks after its outbreak. Bismarck, refusing to be dazzled by the brilliance of the victory, urged the swift conclusion of an honourable peace. Not only did he feel that the preservation of a strong Austria was essential for the maintenance of stability on the Continent, but he also feared that a prolongation of hostilities would enable Napoleon III to intervene in the affairs of Germany. By the preliminary Peace of Nikolsburg (July 26) and the definitive Treaty of Prague (August 23), Francis Joseph was permitted to retain all of his possessions except Venetia, which had been promised to the Italians. There was to be no occupation and only a modest indemnity. The emperor had to acquiesce, however, in the Prussian annexation of Hanover, Nassau, Hesse-Kassel, Schleswig-Holstein, and Frankfurt am Main, in the dissolution of the German Confederation, and in the formation under Hohenzollern leadership of a new federal union north of the Main River. The contest between Berlin and Vienna that had determined the history of central Europe for more than a century was now over.

Bismarck's national policies: the restriction of liberalism. Bismarck's triumph in the military struggle led directly to his victory in the constitutional conflict. Before the outbreak of hostilities he had tried to reach an understanding with the liberal opposition, but the liberals hesitated to make peace with a statesman who had so flagrantly violated the fundamental law of the kingdom. The defeat of Austria changed all that. While the war was still in progress, general elections resulted in important gains for the right. Many voters, elated over the successes of the Prussian armies, expressed their confidence in the government by supporting its adherents at the polls. Some of the ultraconservatives hoped that the Cabinet would now capitalize on its triumph by suspending the constitution and establishing an authoritarian regime. Yet the prime minister recognized that such reactionary schemes would prove futile in the long run. What he wanted was not the suppression of liberalism but an accommodation with it. As soon as peace was concluded, he introduced in the legislature a bill of indemnity granting the government retroactive approval for its conduct of affairs of state without a legal budget. The consequence, as Bismarck had foreseen, was a split in the ranks of his adversaries. Those who argued that there could be no compromise with the principle of constitutional government rejected a reconciliation between crown and parliament based on mutual concessions, but many members of the opposition, who eventually formed the National Liberal Party, decided to accept the settlement offered by the prime minister. Their reasoning was that an obstinate resistance against the Cabinet would only condemn them to sterile dogma-

tism, whereas a willingness to accept what could not be prevented would enable them to influence official policy in the direction of greater freedom. This view prevailed, and on Sept. 3, 1866, the legislature approved the Bill of Indemnity, 230 to 75. By dividing the forces of reform and weakening their sense of purpose, Bismarck won as important a success in domestic affairs as the victory on the field of battle.

His subjugation of liberalism was made possible by the triumph of nationalism. It had long been the prime minister's belief that the achievement of unity could appease the demand for freedom. The success of the Prussian armies provided him with an opportunity to test this assumption. Once Austria had been subdued, he cajoled and bullied the secondary governments above the Main into joining Berlin in the establishment of the North German Confederation. It was the union of a giant and 21 pygmies, for the Hohenzollern kingdom constituted about four-fifths of the territory and population of the confederation. Executive authority was vested in a presidency held in accordance with hereditary right by the rulers of Prussia, who were to exercise the powers of their office with the assistance of a chancellor responsible only to them. The legislature was composed of a federal council, or Bundesrat, whose members were appointed by the state governments, and a lower house, or Reichstag, elected by equal manhood suffrage. Since Berlin had 17 votes out of 43 in the Bundesrat, it could easily control the proceedings with the support of a few of its satellites among the small principalities. Although the Reichstag could theoretically exercise considerable influence over legislation by granting or withholding its approval, parliamentary authority and party initiative were weak and untested, and Bismarck had little difficulty in piecing together a workable majority for his program by a strategy of divide and rule. His support came largely from a combination of moderate liberals and moderate conservatives willing to sacrifice theory for expediency. The federal constitution provided no bill of rights, no ministerial responsibility, and no civilian supervision over military affairs. But it introduced uniformity in currency, weights, measures, commercial practices, industrial laws, and financial regulations. In short, it created the economic unity that the middle class had long demanded and that helped reconcile it to the defeat of its hopes for greater political freedom.

The North German Confederation

Franco-German conflict and the new German Reich. The Seven Weeks' War, by creating a powerful new state in the heart of central Europe, abruptly altered the system of international relations on the Continent. Every government now had to reexamine its diplomatic and military position in the light of the establishment of the North German Confederation. No nation, however, was affected by the victory of the Hohenzollern armies as directly as France. Emperor Napoleon III had encouraged the outbreak of hostilities between Austria and Prussia on the assumption that both combatants would emerge from the struggle exhausted and that the Second Empire of France could then expand eastward against little resistance. The outcome of the war revealed how shortsighted such calculations had been. Instead of profiting from the conflict between Francis Joseph and William I, Paris suddenly confronted a strong and united German state that presented a serious threat to French interests. The imperial government was bound to regard this turn of events with suspicion and hostility. It sought to mitigate its discomfiture by seeking compensation in the Rhineland, Luxembourg, or Belgium. But Berlin succeeded in frustrating these plans, and the conviction began to grow in France that sooner or later a struggle with Germany would be unavoidable. The prospect of a new armed conflict was not unwelcome to Bismarck. He wanted to see national unification consummated by the entry of the southern states into the North German Confederation. Yet public opinion below the Main remained distrustful. Only a common patriotic struggle against foreign aggression might overcome the reluctance of the south to unite politically with the north. Thus in Berlin as well as in Paris there were reasons for seeking a test of strength. The immediate occasion came in the spring of 1870 with the candidacy of Prince Leopold, a relative of William I, for the throne of Spain. The ensuing controversy was cleverly exploited by Bismarck to provoke the French into initiating hostilities which both sides had wanted but for which the French government received the blame.

When France learned of Leopold's acceptance of the offer of the Spanish crown, there were wild protests in Paris and an immediate demand that Leopold be ordered to withdraw. On July 12 Leopold's father renounced the Spanish candidature on his behalf. This was not enough for the French government; it insisted that William I, as head of the Hohenzollern family, should promise that the candidature would never be renewed. This demand was presented to William at Ems by the French ambassador. Though William refused to give a promise, he dismissed the ambassador in a friendly enough way. But when the "Ems telegram," a report of what was happening, reached Bismarck, he shortened it for publication in such a way as to imply that William had refused to see the French ambassador again. The French used this as an excuse to declare war on Prussia on July 19.

Bismarck's calculation that a struggle waged ostensibly against the expansionist designs of Napoleon III would overcome particularism below the Main proved correct. The southern states joined the north in the Franco-German War, and the sense of unity engendered by the brotherhood of arms was soon enhanced by the intoxication of victory. The German troops won one battle after another in hard fighting along the frontier, until on September 2 they forced a large French army, headed by the emperor himself, to surrender at Sedan. The result was the establishment of a republican government in France, which continued to wage the struggle in the name of the old revolutionary ideals of 1793. The generalship of Moltke, however, was too much for the fierce determination of the new regime. Paris capitulated on Jan. 28, 1871, after a long and bitter siege, and on May 10 the Treaty of Frankfurt brought the war officially to a close. The Third Republic had to cede Alsace-Lorraine, pay an indemnity of five billion francs, and accept an army of occupation. It was a Carthaginian peace designed to crush a dangerous rival. The work of national unification in Germany, in the meantime, was successfully completed even before hostilities had ended. Bismarck had entered into negotiations with the southern states soon after the outbreak of war, determined to use patriotic fervour as an instrument for achieving political consolidation. The enthusiasm aroused in central Europe by the victory over France proved too much for the defenders of particularism. On January 18, while Prussian guns bombarded Paris, William I was proclaimed emperor of a united nation at military headquarters in Versailles. The governments below the Main joined the North German Confederation to form a powerful new Reich under the Hohenzollerns. Within a single lifetime, central Europe had completed the transition from cosmopolitanism to nationalism, from serfdom to industrialization, from division to union, from weakness to preponderance, from the Holy Roman Empire to the German Empire. (T.S.H.)

Bismarck's unification of Germany

Germany from 1871 to 1945

THE GERMAN EMPIRE, 1871–1918

The German Empire was founded on Jan. 18, 1871, in the aftermath of three successful wars by the north German state of Prussia. Within a seven-year period Denmark, the Habsburg monarchy, and France were vanquished in short, decisive conflicts. The empire was forged not as the result of an outpouring of nationalist feeling from the masses but from traditional cabinet diplomacy and agreement by the leaders of the states in the North German Confederation with the hereditary rulers of Bavaria, Baden, Hesse, and Württemberg. Prussia, occupying more than three-fifths of the area of Germany and having approximately three-fifths of the population, remained the dominant force in the empire until its demise at the end of another war in 1918.

At its birth Germany occupied an area of 208,825 square miles and had a population of more than 41 million, which

was to grow to 67 million by 1914. The religious makeup was 63 percent Protestant, 36 percent Roman Catholic, and one percent Jewish. The nation was ethnically homogeneous apart from a modest-sized Polish minority and smaller Danish and French populations. Approximately 67 percent lived in villages and the remainder in towns and cities. Literacy was close to universal because of compulsory education laws dating to the 1820s and '30s.

Domestic concerns. From its origins in 1871, the empire was governed under the constitution designed four years earlier by Otto von Bismarck, the Prussian prime minister, for the North German Confederation. This constitution reflected the predominantly rural nature of Germany in 1867 and the authoritarian proclivities of Bismarck, who was a member of the Junker landowning elite. There were two houses: the Reichstag, to represent the people, and the Bundesrat, to represent the 25 states. The former was composed of 397 members elected by universal manhood suffrage and a secret ballot. The constituencies established in 1867 and 1871 were never altered to reflect population shifts, giving rural areas a vastly disproportionate share of power as urbanization progressed. In theory the Reichstag's ability to reject any bill seemed to make it an important reservoir of power; in practice, however, the power of the lower house was circumscribed by the government's reliance on indirect taxes and the Reichstag's limited right to approve the military budget once every seven years. All legislative proposals were submitted to the Bundesrat first and to the Reichstag only if they were approved by the upper house. In addition, the legislative bodies were rarely consulted about foreign policy. Imperial ministers were chosen by and were responsible to the emperor rather than to the legislature. In comparison with the lower houses of the French and British parliamentary systems of the period, the Reichstag was a relatively impotent body.

The Constitution of 1867

A problem that was to plague the empire throughout its existence was the disparity between the Prussian and imperial political systems. In Prussia the lower house was elected under a restricted three-class suffrage system, a propertied franchise that allowed 15 percent of the male population to choose approximately 85 percent of the delegates. A conservative majority was always assured in Prussia, whereas the imperial franchise resulted in increasing majorities for the political centre and left-wing parties. William I (1797–1888) was both German emperor and king of Prussia. Apart from two brief instances the imperial chancellor was simultaneously prime minister of Prussia. Thus, the executives had to seek majorities from two separate legislatures elected by radically different franchises. A further problem was that government ministers were generally selected from the civil service or the military. They often had little experience with parliamentary government or foreign affairs.

The constitution had been designed by Bismarck to give the chancellor and monarch primary decision-making power. Universal suffrage had been proposed because of Bismarck's belief that the rural population would vote for either the Conservative or Free Conservative parties. The Progressives, a left-wing liberal party, were expected to do poorly in the two-thirds of Germany that was rural. Bismarck had not counted on new parties such as the Centre Party, a Roman Catholic confessional party, or the Social Democrats (SPD), both of which began participating in imperial and Prussian elections in the early 1870s. The Centre generally received 20–25 percent of the total vote in all elections. The SPD grew from 2 seats in the first imperial election to 35 by 1890, in which year the SPD actually gained a plurality of votes. Bismarck termed the Centre and SPD along with the Progressives *Reichsfeinde,* or enemies of the empire, because they sought each in its own way to change the fundamental conservative political character of the empire.

Beginning in 1871 he launched the Kulturkampf, a campaign in concert with German liberals against political Catholicism, Bismarck's aim was clearly to destroy the Centre Party. Liberals saw the Roman Catholic church as politically reactionary and feared the appeal of a clerical party to the one-third of Germans who professed Roman Catholicism. Both doubted the loyalty of the Catholic population to the Prussian-centred and, therefore, primarily Protestant nation. In Prussia, the minister of ecclesiastical affairs and education, Adalbert Falk, introduced a series of bills establishing civil marriage, limiting the movement of the clergy, and dissolving religious orders. All church appointments were to be approved by the state. As a result hundreds of parishes and several bishoprics were left without incumbents. Clerical civil servants were purged from the Prussian administration.

The Kulturkampf

The Kulturkampf failed to achieve its goals and, if anything, convinced the Roman Catholic minority that their fear of persecution was real and that a confessional party to represent their interests was essential. By the late



The German Empire, 1871–1918.

1870s Bismarck abandoned the battle as a failure. He now launched a campaign against the SPD in concert with the two conservative parties and many National Liberals. Fearing the potential of the Social Democrats in a rapidly industrializing Germany, Bismarck found a majority to outlaw the party from 1878 to 1890, although it could not constitutionally be forbidden to participate in elections. Party offices and newspapers were closed down and meetings prohibited. Many socialists fled to Switzerland and sought to keep the party alive in exile. During the 1880s Bismarck also sought to win the workers away from socialism by introducing legislation granting them modest pensions, accident insurance, and a national system of medical coverage. Like the Kulturkampf, the campaign against the SPD was a failure, and when the 1890 elections showed enormous gains for the *Reichsfeinde,* Bismarck began to consider having the German princes reconvene, as in 1867, to draw up a new constitution. The new emperor, William II (1859–1941), saw no reason to begin his reign with a potential bloodbath and asked for the 74-year-old chancellor's resignation. Thus, Bismarck, the architect of German unity, left the scene in a humiliating fashion, believing that his creation was fatally flawed. Indeed, his policy of supporting rapid social and economic modernization while seeking to avoid any reform of the authoritarian political system did lead to an atmosphere of persistent crisis.

The economy 1870–90. The empire was founded toward the end of two decades of rapid economic expansion, which saw the German states surpass France in steel production and railway building. By 1914 Germany was an industrial giant second only to the United States. After the establishment of the North German Confederation (1867) the impediments to economic growth were quickly removed. The usury laws and fetters on internal migration disappeared. A uniform currency based on gold was adopted by Bismarck and his National Liberal allies. An imperial state bank was created, and the tough regulations hindering the incorporation of joint stock ventures fell by the wayside. Combined with the euphoria over unification, these changes led to an unprecedented boom between 1870 and 1873. The *Gründerjahre,* as the years after unification were called, saw 857 new companies founded with a capital of 1.4 billion talers—or more new companies and investment in the private sector than in the previous 20 years. Dividends reached an astounding 12.4 percent. The railway system almost doubled in size between 1865 and 1875. Tens of thousands of Germans invested in stock for the first time to demonstrate both their patriotism and their faith in the future of the new German Empire.

The *Gründerjahre*

These halcyon years came to an abrupt end with the onset of a worldwide depression in 1873. The prices for agricultural and industrial goods fell precipitously; for six successive years the net national product declined. A sharp decline in profits and investment opportunities persisted until the mid-1890s. About 20 percent of the recently founded joint stock companies went bankrupt.

In agriculture, the deeply indebted Junker elite now faced severe competition as surplus American and Russian grain flooded the German market. Among the more immediate consequences of the crash was a burst of emigration from the overpopulated provinces of rural Prussia. During the 1870s some 600,000 people departed for North and South America; this number more than doubled in the 1880s. As a result of the depression social and economic questions increasingly preoccupied the Reichstag, while constitutional and political issues were put on the back burner.

It would be incorrect to draw the conclusion that the economy remained in the doldrums for an entire generation. While the 1870s and early 1890s were depressed periods, the 1880s saw significant recovery in industry, if not in agriculture. The British, who paid scant attention to Germany's emergence as an industrial power, began to respect their competitor during this decade.

In adjusting to the "Great Depression," Germany's leaders chose to return to a regulated or administered economy, after a generation of increasingly free trade. The hallmark of the new age was concentration; Germany became the land of big industry, big agriculture, big banks, and big government. The two areas in which the trend toward a controlled economy was most evident were tariff policy and cartelization. Cartel agreements, which were sanctioned by the state, apportioned markets, set standards for manufactured goods, and fixed prices. It is not coincidental that Germany, where the guild system prevailed into the 19th century, should have given birth to the cartel. Cartels arose rapidly in the steel, coal, glass, cement, potash, and chemical industries. Between 1882 and 1895 the total number of business enterprises grew by 4.6 percent, but the number employing more than 50 workers grew by 90 percent.

Cartelization

In 1878–79 Bismarck initiated a significant change in German economic policy, which coincided with his new alliance with the two conservative parties at the expense of the National Liberals. Tariffs were introduced on iron as well as on the major agricultural grains; the latter were raised in 1885 and again in 1887. This departure from liberal economic policy was necessitated by complaints from industrialists, estate owners, and peasants about the terrible impact the depression was having on their respective incomes. Only Britain held out against the protectionist tide that swept Europe in the 1880s. Bismarck's shift, nevertheless, had serious political implications. It signified his opposition to any further evolution in the direction of political democracy. The grain tariffs provided the Junker estate owners of Prussia, who constituted the main opposition to full political emancipation, with subventions that insulated them somewhat from the international market. Thus, the landed elite, major industrialists, the military, and the higher civil service formed an alliance to forestall the rise of social democracy, prevent further political liberalization, and make sure that the uncertainties of the market did not weaken the elites.

Foreign policy 1870–90. Until his resignation in 1890 Bismarck had a relatively free hand in the conduct of foreign policy. After three successful wars he saw his task as promoting peace and gaining time so that a powerful German Empire in the middle of Europe would come to be accepted as natural rather than as an interloper. The Prussian victories had led to great insecurity among the continental powers, who now adopted universal military service in imitation of Germany and maneuvered for defensive alliances so they would not find themselves isolated in the event of war. Bismarck's two areas of concern were the Balkans, where the disintegration of the Turkish Empire could easily lead to conflict between the Habsburg Monarchy and Russia, and France, which nurtured a desire for revenge against the German victors. Each might spark a general European conflagration that would inevitably involve Germany.

Although Bismarck is considered the foremost practitioner of realpolitik, his policies reflected a preference for dealing with monarchies and a disdain for parliamentary governments. In 1873 he negotiated the Three Emperors' League with Russia and Austria-Hungary. The league collapsed in the mid-1870s when rebellion broke out in Turkey's Slavic provinces. In 1877 Russia declared war on Turkey, leading both Britain and Austria-Hungary to express serious concern about Russia's expansionist war aims. When Russia imposed an annexationist peace on Turkey in the Treaty of San Stefano, Bismarck called for an international conference to reconsider the peace treaty and to forestall another military conflict. At the Congress of Berlin in 1878 Bismarck played the role of honest broker among the powers. Russia reluctantly accepted more modest gains and tensions dissipated.

But a conflagration had barely been avoided. Soon after the conference Bismarck negotiated an alliance with the Habsburg Monarchy (1879), which remained in effect through World War I. Thus he tied the fate of the youthful German Empire to the aged multinational empire that faced continuous problems from its many ethnic minorities. The chancellor had clearly opted for the Dual Monarchy over Russia should a war break out. The alliance gave him leverage in Vienna, and he steadfastly used it to prevent a war over the Balkans. His choice of the Habsburg Monarchy resulted from his fear that its dissolution would

The Dual Alliance

lead to Russian hegemony over the Polish, Czech, and other Slavic provinces of the empire. In addition, seven million Austro-German Catholics might seek admission to the German Empire, leading to a strengthening of the hated Centre Party.

Having a solid ally, Bismarck demonstrated his virtuosity by negotiating a revived Three Emperors' League in 1881. He now had influence in St. Petersburg as well as in Vienna to prevent a conflict over the Balkans. In 1882 Italy, fearing French hostility, joined the Dual Alliance with Austria-Hungary, making it into the Triple Alliance. On the surface Bismarck had triumphed. France had no allies for a war of revenge, and the alliance with the Habsburgs and Russia at once gave him influence with the two major adversaries in the Balkans.

The transient nature of this artistry soon became apparent. A crisis in Bulgaria inflamed Russo-Austrian relations, leading to the breakup of the revived league. Once again a war was avoided with Bismarck's intervention, but the Habsburgs and the Romanovs could no longer patch up their differences. Bismarck negotiated a separate Reinsurance Treaty with the tsar in 1887. Nevertheless, France and Russia began courting each other before Bismarck left office.

Between 1870 and 1890 Bismarck earned the respect of European leaders for his pacific policies. Apart from a few colonial additions in the mid-1880s, Germany under his guidance acted as a satiated power. The question remained whether this burgeoning industrial power led by the Junker and industrial elites would continue this policy while the other Western powers carved out world empires.

Leo von Caprivi

Politics 1890–1914. The political structure established by Bismarck in 1867 remained with scant change until the empire's demise in 1918. Leo von Caprivi, Bismarck's successor, was a political neophyte, having spent his entire career in the military. Given the disjuncture between the Prussian and German political systems (see above), Caprivi, surprisingly, sought to work with the parties of the centre and left, Bismarck's *Reichsfeinde.* With their support he reduced the grain tariffs and negotiated long-term trade treaties with Russia, the Dual Monarchy, and Romania. Food prices fell as a result, and industry flourished. National wealth rose rapidly, as did the standard of living of the industrial labour force. The Junker elite were outraged at Caprivi's willingness to sacrifice their interests in behalf of industry and labour. Utilizing their political power in Prussia and their access to the emperor, they were able to force his resignation in 1894, making his chancellorship the shortest before the war. After his resignation, the former general wrote to a friend, "In regard to the 'Junker agrarians' I see only evil, and it appears to me that a revolution by the agrarians is not impossible and for the moment more dangerous than a Social Democratic revolution."

Succeeding chancellors learned from Caprivi's fall that opposition to the landed elite was fraught with peril. Bernhard von Bülow, chancellor from 1900 to 1909, abandoned Caprivi's trade policy and resurrected the alliance of the agrarian and industrial elites.

To many observers the empire was edging toward a monumental crisis as it entered the 20th century. While its authoritarian political system was marked by paralysis, its economy was the most dynamic in Europe. With each election the increasingly urban electorate returned Social Democrats in growing numbers. By 1890 the Social Democrats (who had adopted a Marxist program of revolution at their Erfurt congress in 1891) received more votes than any other party, although four other parties won more seats. By 1912 they had more voters supporting them than the two next-largest parties combined. Both the Centre and Social Democrats were able to create parties with a mass base in German society. The Conservatives, National Liberals, and Progressives were more traditional parties, led by notables who were ill at ease in the world of populist politics. All three were declining, especially the Conservatives, who, despite flirting with anti-Semitism after 1893 by becoming a Christian party, fell to less than 15 percent of the vote by 1912. Many contemporary observers thought that a major crisis was impending between the recalcitrant elites and the increasing number of Germans who desired political emancipation similar to that of Britain and France.

The rise of interest groups

While the Liberals and Conservatives declined in the Reichstag, a new phenomenon on the political scene was the rise of single-issue extraparliamentary interest groups. For the most part organizations such as the Pan-German League, the Navy League, the Farmers League, and the Colonial League were authoritarian in their politics and aggressively expansionist in foreign policy. Their constituencies were overwhelmingly middle-class and educated (except for the Farmers League), and they sought to influence the decision-making process both directly, by impressing the ministers with their strength, and indirectly, by supporting parties that adhered to their goals. Given the wealth and high status of their membership (professors were highly visible as leaders), they were unusually effective in achieving their goals. One of the striking characteristics of the empire was the support it received from the educated strata of the population, despite its elitist constitution.

In the last election during the empire (1912) the Social Democrats scored a great victory, capturing 34.8 percent of the vote and 110 seats. On the local level they had begun to cooperate with the Progressives and occasionally with the Centre Party. Southern states such as Württemberg were moving toward full parliamentary government, and Alsace-Lorraine was given a surprising degree of self-government. Thus, there were some indications that the empire was evolving into a representative democracy. On the other hand, the states of Saxony and Hamburg adopted even more restrictive franchises than Prussia in the years before World War I. Above all Prussia and its Junker, military, and bureaucratic elites, supported by much of the professoriat, stood firm in opposition to any further democratization. Some historians have viewed the outbreak of the war in 1914 as an attempt by these elites to shore up their sagging position by a successful war and annexations, as Bismarck had done in the 1860s when the authoritarian Prussian state was besieged by a liberal opposition.

Economy 1890–1914. The speed of Germany's advance to industrial maturity after 1890 was breathtaking. The years from 1895 to 1907 witnessed a doubling of the number of workers engaged in machine building, from slightly more than one-half million to well over a million. An immediate consequence of expanding industrial employment was a sharp drop in emigration; from an average of 130,000 people a year in the 1880s, the outflow dropped to 20,000 a year in the mid-1890s. The surplus population continued to leave Prussia's eastern provinces, but the destination was the growing and multiplying factories of the Ruhr rather than the Americas. Earlier British fears of German competition now came to fruition. Only one-third of German exports in 1873 had been finished goods; the percentage rose to 63 percent by 1913. Germany came to dominate all the major continental markets except France.

The focus of national wealth as well as population shifted to the urban industrial sector by 1900. Only 40 percent of Germans lived in villages by 1910, a drop from 67 percent at the birth of the empire. Cities of more than 100,000 inhabitants accounted for one-fifth of the population in 1914, compared to one-twentieth at the time of unification. The application of intensive agricultural techniques led to a doubling in the value of all farm products despite a sharp decline in the rural population. Industry accounted for 60 percent of the gross national product in 1913.

The German working class grew rapidly in the Wilhelmian era. Total union membership reached 3.7 million in 1912, of which 2.5 million were affiliated with the socialist unions. Bismarck's welfare legislation covered some 13.2 million workers by 1911. Although German employers were extremely authoritarian and hostile to collective bargaining, the labour force did make significant economic gains. Between 1867 and 1913 the number of hours worked per year declined by 14 percent. Nearly every study of real income shows a rapid rise until 1902 and then a modest increase yearly thereafter. National income

per capita rose from 352 marks to 728 during the life of the empire. Despite these advances industrial workers lacked full political rights, which led a large number of them, including many Roman Catholic workers, to vote for the revolutionary socialist party.

While industrialization was rapid, it occurred only in certain sectors of the economy; other areas were only marginally affected. Some two million Germans persisted in traditional artisanal enterprises as the nation became an industrial colossus. While Germany was characterized by large Junker estates and cartels, it was also the nation of dwarf-sized farms (60 percent of farmers owned less than five acres) and the small workshop. German factories were larger and more modern than their British and French counterparts, but the preindustrial sector was more backward. During depressions those in the traditional trades often turned to anti-Semitism as an ideology that was at once patriotic and anticapitalist.

Foreign policy 1890–1914. Bismarck's successors rapidly abandoned the premises of his foreign policy. The Reinsurance Treaty of 1887 with Russia was dropped, leaving the empire more firmly tied to the Dual Monarchy and Russia free to conclude an alliance with France in 1894. Within four years Friederich von Holstein, a councillor in the political division of the foreign office, had weakened Germany's influence in the Balkans and allowed France to end its isolation. German overtures to Britain remained nugatory.

In 1895 the brilliant young sociologist Max Weber gave an inaugural lecture in Freiburg in which he pointed out that while Germany was establishing a nation-state belatedly the other powers had been founding world empires in Africa and Asia. Weber admonished his audience that Germany had to follow suit or become another Switzerland. Weber's contemporaries needed little convincing.

Weltpolitik

From 1897 to 1912 William II, with his politically astute naval adviser, Alfred von Tirpitz, launched a policy of *Weltpolitik*. Germany's naval power went from being negligible to being second only to Britain's in little more than a decade. Nor did Germany build a navy simply to defend its coastline; rather, the new battleships were capable of challenging the other naval powers on the oceans. Tirpitz was a master of publicity, able to win much of the commercial and industrial middle class to his vision of a mighty empire, whose shipping lanes would be guarded by his fleet. The fleet was also expected to make the monarchy more popular and stem the growth of the left-wing parties. *Weltpolitik* proved to be a colossal failure. Germany came on the imperial scene late, when the choicest territories had already been occupied. Togo, part of New Guinea, and southwest Africa hardly seemed to justify the enormous expenditures on the navy. Moreover, Tirpitz's plans alienated Britain. Germany already had the most powerful army in the world when it fastened on becoming a great naval power. The British found this threatening and negotiated an alliance with Japan in 1902 and another one with France in 1904. In 1907 Britain settled its differences with Russia, and the Triple Entente (including France) was established. Germany now found itself surrounded by three major powers.

The more established powers found Germany meddling everywhere. This was particularly striking, because it followed two decades of Bismarck's policy of avoiding conflict. The Japanese objected to Germany's involvement in China in the 1890s. Russia watched as German power and influence were established in Turkey, its hereditary enemy. The French, of course, still harboured dreams of undoing their defeat in 1870. With Britain also alienated, Bismarck's nightmare of a coalition against the young upstart power, Germany, had become a reality. Bismarck had created a system designed for his extraordinary talents; neither his successors nor the unstable young emperor, William II, shared the master's gifts. Parliament was more or less ignored, although some insightful critiques of *Weltpolitik* were to be heard in the Reichstag. Twice Germany's rulers sought to break up the alliance that was forming against the empire. In 1905 and 1911 Germany created crises over French penetration of Morocco. In each case the British and Russians stood firm, and, even though Germany gained concessions, the Triple Entente remained solid.

In 1912 the chancellor, Theobold Bethmann Hollweg, and his ministers reassessed the decade and a half of *Weltpolitik* and judged it a failure. The army, which had not been expanded since 1894, once again became the pride of the regime. Eastern Europe and the Balkans were now considered the most likely areas for German economic and political penetration. Since Italy had become unreliable, Austria-Hungary was the only ally to be counted on in the event of war. Any threat to the stability of the Dual Monarchy could leave Germany totally isolated.

The assassination of the Austrian heir to the throne, Archduke Francis Ferdinand, by Serbian terrorists in June 1914 augured poorly for the future of Austria-Hungary unless it showed resolve in dealing with the provocation. William II and Bethmann Hollweg urged strong measures against Serbia and reasserted their unconditional loyalty if war should eventuate. With Russia industrializing and the Dual Monarchy increasingly threatened by the aspirations of its minorities, time appeared to be on the side of the Triple Entente. Thus, if war was inevitable, the sooner it came the better. A localized conflict between Austria-Hungary and Serbia with a quick victory for the former would be desirable. If Russia chose to intervene to help its Slavic brother, Serbia, then a general European conflict would ensue. This was acceptable to the German government both because of its pessimism about the long-term strength of the Central Powers (*i.e.,* the German Empire and Austria-Hungary) and because the civilian population could be expected to rally to the war effort if tsarist Russia appeared to bear much of the responsibility. War, which inevitably gives the state increased power over civil society and demands sacrifices for the existing regime, may also have been attractive to elites fearing the growing strength of Germany's socialist movement.

World War I. During the first days of World War I Germans experienced a sense of bonding that had eluded them since the founding of the empire. Differences of class, religion, and politics were transcended as Germans flocked to their city centres to show their enthusiastic support for the impending conflict. Overwhelmingly, the parties, including the Social Democrats, voted for war credits. The euphoria of the early days masked Germany's dangerous situation. The Triple Entente commanded the seas, outnumbered Germany in population by three to one, and had access to the world's natural resources through their empires and close contact with the United States. Germany was immediately blockaded and had to rely on its own resources and those of contiguous nonbelligerents like The Netherlands, Denmark, and Switzerland. The Central Powers did have interior lines of transit, which was valuable in a two-front war. They also had a unified command structure—in contrast to the Triple Entente, where rivalries resulted in three different wars being fought simultaneously with little coordination.

The Schlieffen Plan

General Alfred von Schlieffen, chief of the German general staff, had recognized Germany's vulnerability in a two-front war and saw the best hope in an overwhelming attack against France through Belgium. If all went according to plan, France's industrial region would be occupied in six to eight weeks and Paris itself surrounded. The slow-moving Russians would occupy Prussia's eastern rural provinces, facing only a modest-sized German military force. After France's concession of defeat following the occupation of its capital, whole armies would be shifted to the eastern front to drive the Russians out. Schlieffen died in 1913, and the plan was put in motion by General Helmuth von Moltke. As often happens in history, the plans of men may go awry in ironic ways. The western armies of Germany did, indeed, move through neutral Belgium but were stopped at the Battle of the Marne (September 1914) in northern France. Meanwhile, Paul von Hindenburg was reactivated at age 67 and sent with Major General Erich Ludendorff to halt the Russian advance into East Prussia. There the Germans unexpectedly defeated two large Russian armies at the Battle of Tannenberg (August 1914).

The western front turned into a war of attrition as the two sides built opposing trenches from the Swiss border to the English Channel. For three and a half years neither

side moved more than 30 miles despite titanic battles at Verdun, the Somme, and Ypres. In the east the outmanned German forces, with the help of the Austrians, inflicted a series of costly military defeats on Russia, but, given the vast plains of eastern Europe, the Central Powers were unable to knock Russia out of the war until after the seizure of power by Lenin in the October Revolution of 1917. In 1916 Ludendorff and Hindenburg became joint heads of all German land forces and recognized, as had their predecessor Erich von Falkenhayn, that the war would be won or lost on the western front. With Italy (1915) and Romania (1916) entering the war on the side of the Triple Entente, the Central Powers faced an almost impossible situation in a war of attrition. The two generals became de-facto rulers of Germany and sought the mobilization of the whole society for total war. More than 11 million men, some 18 percent of the population, were in uniform, of whom almost two million were ultimately killed. Germany was unable to feed itself, and after the severe winter of 1916–17 malnutrition and even starvation were not uncommon.

On the diplomatic front the elites ruling Germany planned for vast annexations of Russian, Belgian, and French territory as well as for an African empire. The war costs were to be paid by the defeated powers of the Triple Entente. At no time during the war did the German government engage in serious negotiations to restore the sovereignty of Belgium or to return to the status quo ante. Nor were the allies very interested in a negotiated peace, but their situation was not as desperate as Germany's.

Ludendorff and Hindenburg adopted an all-or-nothing policy in regard to victory. They created an independent state of Poland in 1916, which prevented serious negotiations with Russia for a separate peace. They adopted submarine warfare in 1917, despite the knowledge that it would bring the United States into the war, because it offered a slim hope of quick victory if Triple Entente ships carrying men and supplies could be prevented from reaching France. Ludendorff also mounted a major offensive in April 1918, ignoring Woodrow Wilson's proposal of Fourteen Points for a future peace and failing to offer any peace terms of his own. When asked what would happen if the offensive failed, he replied, "Then, Germany will be destroyed."

In 1917 the Reichstag, following the lead of the Centre Party, passed a peace resolution based on no annexations. Social Democrats and Progressives rallied to support the resolution. The military and civilian leadership ignored the resolution and enforced a draconian peace on Russia and Romania in 1917–18. When the major battle in the west was brewing in April 1918, there were more than a million soldiers in the east to enforce the Treaty of Brest Litovsk with Russia.

Clearly, the military, agrarian, and industrial elites who ruled Germany viewed themselves as involved in two wars simultaneously, one against the Triple Entente and the other against the aspirations of the German people for full political emancipation. The latter conflict dictated victory at all costs on the military front. Defeat or a compromise peace on the battlefield would inevitably lead to democratization because of the sacrifices endured by many millions of workers, farmers, and artisans. In November 1914 Alfred Hugenberg, the major industrialist and subsequent ally of Hitler, told German entrepreneurs, "The consequences of the war will be unfavourable to employers and industry in many ways. One will probably have to count on a very increased sense of power on the part of the workers and labour unions, which will find expression in increased demands on the employer for legislation. It would, therefore, be well advised in order to avoid internal difficulties to distract the attention of the people and to give fantasies concerning the extension of German territory room to play." William II felt compelled to promise an eventual end to the restrictive Prussian franchise in his Easter message of 1917. Shortly thereafter the Fatherland Party was established with enormous support from the elites. Its program included a Carthaginian peace and maintenance of the Prusso-German political system.

The Ludendorff offensive of April 1918 made great breakthroughs in the west. But the effects of four years of attrition were apparent. The military did not have the reserves to take advantage of the initial gains. With almost a million fresh American troops in France, the Allies launched a counterattack that quickly gave them the initiative. Slowly the German forces began retreating. On August 8 the German army suffered a severe defeat in northern France, and not long thereafter William II installed a new, more liberal, government in Berlin, headed by Prince Max von Baden. The new ministers were informed that the war was virtually lost, and they were advised to seek an immediate armistice. Before the negotiations were successful, revolution broke out in the German navy and spread to the military and urban workers. William II was forced to abdicate as emperor of Germany and king of Prussia. The Social Democrats took power at this appalling moment of defeat, while the former military and civilian leaders sought to escape responsibility for the calamity. A civilian, Matthias Erzberger of the Centre Party, signed the armistice, which took effect Nov. 11, 1918. (For further discussion of World War I, see the article WORLD WARS, THE.) (Ke.B.)

THE GERMAN REPUBLIC, 1918–33

The republic proclaimed early in the afternoon of Saturday, Nov. 9, 1918, is often called the "accidental republic." When Friedrich Ebert, the leader of the so-called Majority Socialists, accepted the imperial chancellorship from Maximilian, Prince of Baden, it was with the understanding that he would do his utmost to save the imperial system from revolution. Ebert believed that the only way to accomplish this would be by transforming Germany into a constitutional monarchy. Elections would have to be held for a constituent assembly, whose task it would be to draw up a new constitution.

Defeat of revolutionaries, 1918–19. Ebert, however, was faced with a precarious situation. The dangers confronting him were mounting all over the country. Four and a half years of war and sacrifice were giving way to a war-weariness that was leaving the imperial system, as well as its emperor, discredited. Shortages of food and fuel had rendered the population vulnerable to the influenza epidemic sweeping Europe. On October 18 alone Berlin authorities had reported 1,700 influenza deaths. Independent Socialists in Munich had forced the abdication on November 8 of Bavaria's King Ludwig III and proclaimed a Bavarian socialist republic. The port cities along the North Sea and the Baltic Sea were falling into the hands of sailors' and workers' and soldiers' councils (*Räte*) in the wake of the naval mutiny at Kiel in early November. Berlin's radical leaders, Karl Liebknecht and Rosa Luxemburg, were eager to transform Germany into a republic of workers' and soldiers' councils (a *Räterepublik*) in imitation of the soviet republic being established by the Bolshevik leaders in Russia. As Ebert was accepting the reins of government in the Reichstag building on November 9, Liebknecht was addressing a rally of his own followers in front of the deserted Royal Palace about a mile away. Among Marxist revolutionaries the view was widespread that the Bolshevik revolution was merely the spark that would set off the worldwide proletarian revolution predicted by Karl Marx. Inevitably, that revolution would have to spread to Germany. Given this ideologically charged scenario, Liebknecht confidently anticipated his destiny to become the German Lenin.

While the Liebknecht rally was proceeding in front of the Royal Palace, an angry crowd was gathering before the governmental headquarters in the Reichstag building. Because Ebert had just left the building, his friend and fellow Majority Socialist Philipp Scheidemann felt called upon to address the crowd. To meet its inevitable demands for change and to forestall whatever Liebknecht might be telling his followers, Scheidemann in his speech used the phrase "Long live the German Republic!" Once said, the proclamation of a republic could not be withdrawn. Ebert was furious when he learned of Scheidemann's "accidental" proclamation, but he realized that there was no turning back. He spent the afternoon seeking partners to form a provisional government to run the newly pro-

claimed republic. By nightfall he managed to persuade the Independent Socialists, a party that in 1917 had split from the Majority Socialists over the continuation of the war, to provide three members of a provisional government. To gain their cooperation, Ebert had to agree to naming the provisional government the Council of Peoples' Commissars and to transforming Germany into a vaguely defined social republic. This promise notwithstanding, Ebert still hoped that elections to a constituent assembly would lead to the creation of a moderate democratic republic. The Independent Socialists, however, though not as radical as Liebknecht, held to their vision of a socialist *Räterepublik*. They hoped that workers and soldiers would elect a multitude of councils across the entire nation during the next weeks, which they assumed would establish the foundation for a genuinely socialist republic.

For the time being, however, Majority and Independent Socialists were joined in providing provisional governance for the defeated German nation, which everywhere seemed on the verge of collapse. Although the armistice of November 11 ended the fighting, it did not end the Allied blockade. The winter of 1918–19 brought no relief in the shortages of food and fuel, and the flu epidemic showed no signs of abatement. Soldiers returning from the military fronts by the hundreds of thousands were left stranded, jobless, hungry, and bitter—grist for the mill of revolution.

The push for revolution, led by an enthusiastic Liebknecht and a more reluctant Luxemburg, came on Jan. 6, 1919, encouraged by Soviet Russia and further prompted by fear that Ebert's plans for the election of a constituent assembly, scheduled for January 19, might stabilize the German situation. The Spartacists, now officially the Communist Party of Germany, initiated massive demonstrations in Berlin and quickly seized key government and communications centres.

The Spartacist revolt

The events of "Spartacist Week," as the radical attempt at revolution came to be known, demonstrated that Germany was not nearly as ripe for revolution as its leaders had thought it to be. As Luxemburg had feared, mass support among German workers for communism did not exist; most of them remained loyal to the Independent Socialists or to Ebert's more moderate and democratic vision of socialism. The German Army, moreover, had recovered its nerve and was determined to prevent a further move to the left. In December the army had begun secretly to train volunteer units drawn from the sea of soldiers returning from the front. These so-called Freikorps ("free corps") units became the basis for the dozens of small right-wing armies that during the next years roamed the country, looking for revolutionary activity to suppress. The Spartacist revolt, which was confined largely to Berlin, was put down within a week by some 3,000 Freikorps men. When Liebknecht and Luxemburg were captured on January 15, they were both shot at the initiative of Freikorps officers. Although sporadic revolutionary activity continued elsewhere in Germany during the next months, its failure in Berlin clearly marked its doom. The proclamation on April 4, 1919, of a *Räterepublik* in Bavaria revived radical fortunes only briefly; Freikorps units put down the radical Bavarian republic by the end of the month.

The collapse of the Spartacist revolt greatly enhanced the chances for Ebert's vision of Germany's future to prevail. Moreover, the meeting of a national congress of workers' and soldiers' councils in mid-December 1918, upon which the Independent Socialists had pinned their own hopes for creating a socialist republic, proved to be far less radical than expected; it did nothing to interfere with Ebert's plans to elect an assembly to draw up a democratic constitution. The elections on Jan. 19, 1919, which for the first time included women, produced a resounding victory for Ebert's conception of democracy. Three of every four voters gave their support to political parties that favoured turning Germany into a democracy. After months of turmoil Germany was to become a democratic republic. The assembly began its deliberations on Feb. 6, 1919, choosing to meet in Weimar, where it believed itself less vulnerable to radical political interference than in Berlin.

On Jan. 18, 1919, representatives of the powers victorious over Germany began the deliberations in Paris that would establish a European peace settlement. Germany's new democratic leaders placed high hopes in the prospects for this settlement. Woodrow Wilson's Fourteen Points seemed to promise Germans national self-determination as well as to encourage the efforts to transform Germany into a democracy. When the German constituent assembly met in Weimar for the first time on February 6, it immediately declared itself sovereign over all Germany. It selected a provisional government—with Ebert as president and Scheidemann as chancellor—whose first major task was to prepare for the expected invitation to Paris to negotiate a treaty of peace with the empire's former enemies.

But the invitation for a German delegation to come to Paris did not arrive until early April. Rather than being treated as a fellow—if fledgling—democracy, Germans soon learned that they were still viewed as the "bad boys" of Europe. Wilson's idealism had been forced to yield to still-fresh wartime resentments being articulated by the leaders of the French, British, and Italian delegations. There were to be no negotiations about a peace treaty. Germany would simply be handed a treaty and told to sign.

The Treaty of Versailles. The treaty in its final form contained many provisions that the Germans had fully expected. That Alsace-Lorraine was to be handed back to France was no surprise; nor were the small territorial adjustments along the border with Belgium. The plebiscite allowing the Danish population of northern Schleswig to choose between joining Denmark or remaining with Germany was unarguably consistent with the principle of national self-determination. But this principle, the Germans expected, would also justify a union between Germany and the Germans of what now remained of Austria after the collapse of the previous November. More serious to Germany was the stipulation that its coal-rich Saar region was to be taken over by the League of Nations and the coal given to France to aid its postwar reconstruction. Eventually a plebiscite was to allow Saarlanders to choose whether or not they wished to be returned to Germany.

On its eastern frontier Germany was forced to cede to the newly restored Poland the province of West Prussia, thereby granting Poland access to the Baltic Sea while losing land access to the province of East Prussia. Danzig was declared a Free City under the permanent governance of the League of Nations. Much of the province of Posen, which, like West Prussia, had been acquired by Prussia in the late 18th-century partitions of Poland, was likewise granted to the restored Polish state. Also lost to Germany as the result of a later plebiscite was a significant portion of coal-rich Upper Silesia.

Overseas, Germany was forced to yield control of its colonies. Although these colonies had proven to be economic liabilities, they had also been symbols of the world-power status Germany had gained in the 1880s and '90s. More immediately serious were the treaty's commercial clauses that took from Germany most of its foreign financial holdings and reduced its merchant carrier fleet to roughly one-tenth of its prewar size.

The treaty's provisions for disarming Germany were to be, the Allied leaders promised, merely the first step in a worldwide process of disarmament. To insure that Germany would not revive as a military power, its army was to be reduced to 100,000 men and would not be allowed to produce tanks, poison gas, or military planes. Moreover, Germany's frontier with France was to be permanently demilitarized; German military forces were to remain behind a line 31 miles east of the Rhine. The treaty also called for the dissolution of the German General Staff, the military command structure of the German Army, which the Allies believed to be the engine of German aggression. The navy, too, was to be dismantled and limited to 15,000 men, a half-dozen battleships, and 30 smaller ships. A prohibition on the building of submarines was absolute. Germany's compliance with the treaty's terms was to be assured by an Allied occupation of the Rhineland and the presence of the Inter-Allied Commissions of Control.

The most resented terms of the treaty, however, were the

so-called honour clauses: Articles 227 through 230 gave the Allies the right to try individual Germans, including the former emperor, as war criminals; Article 231, often called the war-guilt clause, provided the justification for Article 232, setting up a commission to collect reparation payments, the total of which was eventually set at 132 billion goldmarks (about 32 billion gold dollars). German bitterness over these honour clauses was nearly universal. Almost no German believed his country responsible for the outbreak of war in 1914. Technically, Article 231 did not declare Germany to be solely responsible for causing the war, merely "for causing all the loss and damage" suffered by the Allies in the war "imposed upon them by the aggression of Germany and her allies." Germans read it as an accusation of guilt, however, and understood it as the cynical product of victors' justice.

The German provisional government in Weimar was thrown into upheaval when it learned of the treaty's full terms. "What hand would not wither that binds itself and us in these fetters?" Scheidemann asked, and he resigned rather than accept the treaty. Army chief Paul von Hindenburg did the same, after declaring the army unable to resume the war under any circumstances. Only an ultimatum from the Allies finally brought a German delegation to Paris to sign the treaty on June 28, 1919, five years to the day since the assassination of Archduke Francis Ferdinand.

The Weimar Constitution. During the month following the signing of the treaty, the Weimar constituent assembly completed its drafting of a constitution for the new republic. The result was widely hailed as the most modern democratic constitution of its day. The document provided for a popularly elected president who was given considerable power over foreign policy and the armed forces. In Article 48 he was also given emergency decree powers to protect the republic from crises initiated by its opponents from either the left or the right. The president was empowered to nominate the chancellor, whose government required the confidence of the lower house of the parliament, the Reichstag, elected by universal suffrage and a system of proportional representation. An upper house, the Reichsrat, was composed of delegates appointed by the governments of the federal states, the *Länder*.

The most modern features of the Weimar constitution, the provisions for popular initiative and referendum, were designed to enable the electorate on its own initiative, by way of petition, to introduce bills into the Reichstag and to force the body to a vote on the proposal. If the bill was voted down, a national referendum was to be held to allow the electorate to pass the bill into law against the wishes of the Reichstag. Never again could a government afford to ignore the wishes of the voters.

The Weimar constitution was promulgated formally on Aug. 11, 1919, ending the provisional status of government in Germany that had begun with Scheidemann's proclamation of a republic the previous November. In September the government, judging the situation sufficiently safe in Berlin, returned to the capital. But it did not yet consider it sufficiently safe to risk the nationwide elections for president or for a Reichstag to replace the constituent assembly. Instead the assembly prolonged Ebert's provisional term as president for three years; elections for the Reichstag were delayed until June 1920.

Years of crisis, 1920–23. In its early years the new German democracy was beset by continuing turmoil. The Treaty of Versailles, quickly labeled "the *Diktat*" by the German public, galvanized the resentment that had accumulated during the war, much of which was turned back on the republic itself. Its enemies began to put the blame for the hated treaty upon the republic's socialist and democratic progenitors, whom they accused of having undermined Germany's efforts in the final stages of the war. A revived and radicalized right wing asked whether the German Army might not have been stabbed in the back by traitors on the home front. Racist circles took seriously the notorious *Protocols of the Learned Elders of Zion*, a fraudulent document fabricated in Russia in 1895 and published in Germany in 1920. It allowed the conclusion that all of recent history, including World War I, was the product of a conspiracy of Jews seeking to control the world. Roving Freikorps units contributed to the brutalization of German political life. In March 1920 one of these units, under the command of the former naval captain Hermann Ehrhardt, succeeded in briefly establishing its control of the government in Berlin. This so-called Kapp Putsch, named after the conservative politician Wolfgang Kapp, who had planned it, was put down not by the army but by a general strike of Berlin's socialist and communist workers. A similar right-wing putsch in Bavaria was successful, however, and from that point forward radical groups of the right found protection and a degree of nurture in this south German state. By the end of 1922 there had been nearly 400 political assassinations, the vast majority of them traceable to the right wing. The victims included prominent politicians such as Matthias Erzberger, a signatory to the Armistice of 1918, and Walther Rathenau, the foreign minister.

Kapp Putsch

The difficulties in which the new democracy found itself were already reflected in the June 1920 elections to the first Reichstag. The Weimar coalition parties, the SPD (socialist), the Centre Party (Roman Catholic), and the Democrats, which in January 1919 had received over 75 percent of the vote, this time managed to win only 43.5 percent. Contributing to the problems Weimar faced in the early 1920s was the escalating rate of inflation that was eventually to destroy the German mark. Although the inflation was rooted in the huge debt that Germany had amassed in financing the world war, the hyperinflation of 1923 was triggered by the French-Belgian military occupation in January 1923 of the German industrial district in the Ruhr valley. The occupation occurred in retaliation for Germany's having fallen behind in its reparation payments and was intended to force German industry to provide compensation for the French and Belgian losses. Rather than accede quietly to the humiliation of occupation, the German government urged workers and employers to close down the factories. Idle workers were paid for the next months with a currency inflating so rapidly that printers gave up trying to print numbers on bills. By mid-1923 the German mark was losing value by the minute. A loaf of bread that cost 20,000 marks in the morning would cost 5,000,000 marks by nightfall. Restaurant prices went up while customers were eating. Workers were paid twice a day. On November 15, when the collapse came, it took 4.2 trillion German marks to buy a single American dollar.

The hyper-inflation of 1923

The social and political cost of the hyperinflation was high. Scholars note that the inflation did more to undermine the middle classes than the presumably socialist revolution of 1918. A lifetime of savings would no longer buy a subway ticket. Pensions planned for a lifetime were wiped out completely. Politically, the hyperinflation fueled radicalism on both the left and the right. The Communists, badly set back by their failure in January 1919, believed the prospects for a successful revolution to be greatly improved. In Munich the leader of the small National Socialist German Workers' Party, Adolf Hitler, used the turmoil to fashion an alliance with other right-wing groups and attempt a coup in November 1923 that sought to use Bavaria as a base for a nationalist march on Berlin. His hope was to overthrow the democratic system of Weimar that he believed was responsible for Germany's political and economic humiliation. Neither the radicals of the right nor those of the left succeeded in imposing their will. In the short run they did not succeed because of ineptitude and miscalculation; in the long run they failed because the government sponsored a currency reform that led to a restabilization of the mark and also decided to end its policy of passive resistance in the Ruhr in exchange for an end to the occupation and a rescheduling of the reparation payments it owed to the Allies.

The Weimar Renaissance. Amid the political and economic turmoil of the early 1920s Germany's cultural and intellectual life was flowering. The so-called Weimar Renaissance brought to Germany the fulfillment of the Modernist revolution, which in the late 19th century had begun to transform the European aesthetic sensibility. The Modernist rejection of tradition seemed to suit perfectly the need of many Germans for new meanings and values

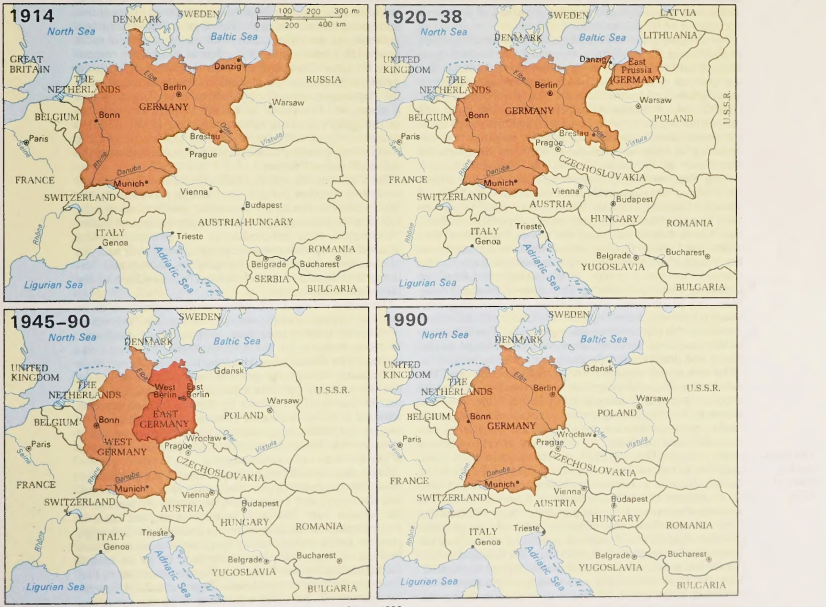
to replace those destroyed by the war. "A world has been destroyed; we must seek a radical solution," said the young architect Walter Gropius upon his return from the front in late 1918. In 1919 Gropius became the founder and first director of the Bauhaus school of design in Weimar, the most important institution in Germany for the expression of Modernism's aesthetic and cultural vision. Bauhaus artists believed themselves to be creating a new world through their painting, poetry, musical compositions, theatrical productions, and architectural constructions. The legacy of German Modernism in general, and of the Bauhaus in particular, is most immediately evident in the stark steel-and-glass high-rise buildings whose clear and clean lines have come to dominate the skylines of the world's cities. Moreover, the paintings and the sculptures decorating them, as well as the design of the furniture and the lighting fixtures, are heavily influenced by the aesthetic principles articulated in Weimar during the 1920s.

Beyond the Bauhaus, painters like George Grosz, Max Beckmann, and Otto Dix pursued an artistic objective described as Expressionism; *i.e.*, they were interested in depicting their emotional responses to reality rather than reality itself. In music the rejection of tonality by composers such as Arnold Schoenberg, Anton von Webern, and Alban Berg broke a centuries-old tradition in music. At the juncture between popular and serious music the composer Kurt Weill collaborated with the poet Bertolt Brecht to create in 1928 *The Threepenny Opera*, a bitterly satiric musical play in which the world of modern capitalism was equated with that of underworld gangsterism. In films such as *The Cabinet of Dr. Caligari*, distorted sets and unusual camera angles probed for disturbing truths behind the surface appearances of reality.

Not everyone was pleased by the Modernist attack on tradition. Opera performances and theatrical productions were often interrupted by angry audiences. Siegfried Wagner, the son of the composer Richard Wagner, deplored a Modernist version of his father's *The Flying Dutchman*, calling the production an example of "cultural bolshevism." Other artists, the novelist Thomas Mann, for example, chose to remain above the fray in the Olympian heights of German *Kultur*, an effort for which Mann in 1929 won the Nobel Prize for Literature.

Years of economic and political stabilization. The financial recovery that began with the restabilization of the German currency in late 1923 was undergirded in 1924 when the Allies agreed to end their occupation of the Ruhr and to grant the German government a more realistic payment schedule on reparations. A committee of the Allied Reparations Commission headed by the American financier and soon-to-be vice president Charles Dawes had recommended these changes as well as urged the Allies to grant sizable loans to Germany to assist its economic recovery. The Dawes Plan marked a significant step in the upswing of the German economy that lasted until the onset of the Great Depression. The 800 million gold marks in foreign loans had by 1927 enabled German industrial production to regain its 1913 prewar high. That same year the Reichstag contributed to the vital need for social and class reconciliation by voting for a compulsory unemployment insurance plan. Reconciliation on the political level seemed achieved in 1925 when the 77-year-old Hindenburg was elected to succeed the deceased Friedrich Ebert as president. Although no democrat, the aged field marshal took seriously his duty to support the constitution and the republic.

The Dawes Plan



Changes in Germany's territory and internal status from 1914 to 1990.

The guiding spirit in German foreign policy from 1924 through 1929 was the foreign minister, Gustav Stresemann, who firmly believed that Germany was more likely to gain relief from the harshness of Versailles by trying to fulfill its terms than by stubbornly continuing to resist them. Stresemann's efforts ushered in what came to be known as "the era of fulfillment." It began in December 1925 when Germany signed the Locarno treaties, in which it guaranteed to maintain the new postwar boundaries with France and Belgium and to submit to international arbitration any boundary disputes that might arise in the east with Poland or Czechoslovakia. Germany was formally readmitted into the family of nations by being granted membership in the League of Nations in September 1926. In 1928 Germany became party to the most dramatic symbolic gesture of postwar reconciliation, the Kellogg-Briand Pact, which promised to outlaw the resort to aggressive war; this agreement was signed by nearly all the world's major countries during the next year.

The economic and political stabilization of the German republic seemed to be reflected in the May 1928 elections to the Reichstag. The antirepublican parties of the left and right together received only 13 percent of the total vote, with the Communists receiving 10.6 percent and the Nazis taking only 2.6 percent. Germany's reintegration into the international political structure was reflected also in the decision in early 1929 by the Allied Reparations Commission to make a final disposition of the reparations question. The committee appointed to make recommendations in this matter was headed by Owen D. Young, an American business executive. The Young Committee proposed that German reparations be reduced to about 37 billion gold marks, less than one-third of the 1921 total, and that payments be stretched until 1988. It also called for the dissolution of the Reparations Commission and for an immediate end to what remained of the Allied occupation of the Rhineland.

The German government, seeing the obvious advantages in the Young Plan, officially accepted its terms in August 1929. Right-wing parties of the opposition, however, saw the plan as nothing less than a renewal of Germany's humiliation. Led by the German National Peoples' Party (DNVP) and its leader Alfred Hugenberg, the press and movie-industry lord, the nationalist opposition seized upon the constitutional processes for popular initiative and referendum in order to force the government to reverse its acceptance of the plan. To run the opposition's anti-Young Plan campaign, Hugenberg engaged Adolf Hitler, the leader of the apparently moribund NSDAP (National Socialist German Workers' Party). The objective was to force the German government into a repudiation of the reparations debt as well as of the war-guilt clause of Versailles upon which the debt rested. German signatories to the Young Plan, moreover, were to become liable to the charge of treason. The right wing's initiative did force the Reichstag into reconsidering its approval of the Young Plan, but to no avail. The national plebiscite that necessarily followed found only 13.8 percent of the voters favouring the objectives of the right wing. The bitterness of the campaign, however, may have contributed to the illness and death during the campaign of Stresemann.

The end of the republic. An unintended effect of the anti-Young Plan campaign was to give widespread public exposure to the little-known Adolf Hitler. Hitler used his access to the Hugenberg-owned press empire and to its weekly movie newsreels to give himself and his Nazi movement national publicity. An additional assist to Hitler's career came on Oct. 29, 1929, with the stock market crash on Wall Street, an event that signaled the onset of what quickly became a worldwide depression. The crash had an immediate effect in Germany as American investors, anxious about their financial position, began withdrawing their loans to Germany. German indebtedness to these investors had by 1929 reached nearly 15 billion marks. Prices on the German stock exchanges fell drastically during the last month of the year. Business failures multiplied. Early in 1930 Germany's second-largest insurance firm collapsed. Unemployment rose to three million during the course of the year. By the winter of 1932 it reached six million. Germany's industry was working at no more than 50 percent of its capacity, and the volume of German foreign trade fell by two-thirds between 1929 and 1932.

The stock market crash of 1929

The first critically important political effect of the economic crisis came in March 1930 when the government coalition fell apart over the rising cost of maintaining the unemployment program adopted in 1927. The SPD, representing labour, and the Peoples' Party (VP), representing business, were unable to agree on the size of the government's contribution to the fund, and therewith the government fell. When a new coalition could not be found, parliamentary democracy in Germany came to an end.

President Hindenburg was forced by the situation to invoke his emergency powers (Article 48), which he used to appoint Heinrich Brüning of the Catholic Centre Party as chancellor. For the next two years, until May 30, 1932, Brüning governed without a parliamentary majority, deriving his authority from the powers residing in the office of President Hindenburg. However well-intentioned, Brüning's deflationary economic policies were unable to stem the tide either of the depression or of its social and political ravages. His fateful decision to call for Reichstag elections in September 1930, moreover, inadvertently opened the door for the enemies of Weimar democracy. Together the Nazis and the Communists gained nearly one of every three votes cast. The 18 percent acquired by the Nazis represented seven times the total number of votes they had received in 1928; the 13 percent won by the Communists represented a healthy gain over the 10 percent of 1928.

Heinrich Brüning

Although bitterly opposed to each other, the Nazis and Communists during the next two years succeeded in mobilizing the political and economic resentments generated by the depression. Hitler's charismatic appeal and the youthful energies of his movement were attractive to large segments of a populace fearful of being ruined by economic and social disaster. Hitler's record as a war veteran lent authority to the hypernationalism he expressed in racist terms. His identification of the Jew as the enemy responsible for all of Germany's ills, be it the defeat of 1918, the Treaty of Versailles, the reparations, the inflation, or now the depression, seemed plausible to many eager to find a scapegoat. The power of Hitler's appeal was reflected in the party's growing membership lists (1929: 170,000 members; 1932: 1,378,000 members) and in the swelling ranks of the party's paramilitary SA (Sturmabteilung), the infamous storm troopers.

The Communists, unlike Hitler, found it difficult to extend their support beyond the German working classes. Their political flexibility, moreover, was limited by the fact that they increasingly fell under the control of Stalin. Nonetheless, their self-confidence was greatly enhanced because the depression seemingly confirmed their analysis of the inevitable collapse of capitalism. Hitler and National Socialism they perceived merely as products of the last phase of capitalism.

The winter of 1931–32 witnessed the depression at its depths. Unemployment was still rising; the succession of business failures resembled rows of falling dominoes. Brüning, helpless in the face of these problems, was dubbed "the hunger chancellor" by his critics. Some hope of breaking through the political impasse was offered by the series of critical state legislative elections scheduled for the spring of 1932 and by the presidential election required at the expiration of Hindenburg's first term. The 84-year-old Hindenburg was with great difficulty prevailed upon to seek a second term. The year 1932 was to be one of continuous election campaigning.

Hitler chose to challenge Hindenburg for the presidency. Although Hindenburg was eventually reelected, a runoff was necessary, and Hitler was able to take 37 percent of the popular vote. His larger aim, however, had been to make himself the leading, or only, candidate for Brüning's position as chancellor. Hindenburg did choose to replace Brüning in May 1932, but rather than Hitler he named the political dilettante Franz von Papen. Desperate to find for himself a base in parliament, Papen called for Reichstag elections in July. The result was a disaster for Papen and another triumph for the Nazis, who took 37 percent

of the vote, the largest total they were ever to acquire in a free election. The Communists won 15 percent of the vote. Thus the two parties dedicated to destroying German democracy held a majority in the Reichstag. Still, Hitler did not get the chancellorship. In November Papen called for another Reichstag election in the hope of gaining parliamentary backing. Again he failed, although the Nazi vote fell by 4 percent. The Communist vote, on the other hand, rose to nearly 17 percent. In early December, when Hindenburg decided to replace Papen, he again ignored Hitler, choosing instead a friend from the army, General Kurt von Schleicher.

In the Nazi camp there was bitter frustration at the end of 1932. The party was deeply in debt and demoralized by the year's endless campaigning. Putschist elements in the party, never persuaded that elections could bring the party to power, were growing increasingly restive. So deep was the frustration that on December 7 Gregor Strasser, second only to Hitler in the party, broke with the Nazis and retired from politics.

THE THIRD REICH, 1933–45

The Nazi revolution. When Hitler finally became chancellor on Jan. 30, 1933, it was not on the crest of a wave of popular support but as the result of backroom political intrigue by Schleicher, Papen, and the president's son, Oskar von Hindenburg. Only Hitler, they believed, could bring together a coalition with Hugenberg's DNVP and possibly the Centre Party that could command a majority in the Reichstag. They assured the reluctant president that Hitler's radical tendencies would be checked by the fact that Papen would hold the vice-chancellorship and that other conservatives would control the crucial ministries, such as those of war, foreign affairs, and economics. The Nazis themselves were restricted to holding the chancellorship and the powerless ministry of the interior. As a sop to the Nazis, Hermann Göring was granted ministerial status but given no portfolio; yet, significantly, he became interior minister in the state of Prussia, which gave him control over the largest police force in Germany.

The Nazis brought with them an ideology that purported to champion the common man, whom they portrayed as victimized in a world controlled by Jews. Anti-Semitism and notions of German racial superiority were at the core of this ideology, which, in its particulars, was also a catalog of resentments that had accumulated in German society since November 1918. Heading the list were the humiliations associated with Versailles, but not far behind were resentments of big business, big banks, big department stores, and big labour, as well as resentments of the divisiveness and inefficiencies seemingly being fostered by political parties.

Neither the 25-point party program (1920) nor Hitler's autobiographical *Mein Kampf* (1925) contained clear conceptions of how the German world would be structured under the Nazis, but Hitler and his propagandists had communicated clearly that the changes would be fundamental and come at the expense of Germany's racial enemies. Racially superior Germans were to be gathered into a tightly knit *Volksgemeinschaft*, or racial community, in which divisions of party and class would be transcended in a spirit of racial harmony, a harmony that would necessarily exclude people of inferior blood. The logic of this construct required a solution to what the Nazis called "the Jewish problem." At the very least it called for a reversal of the trend, more than a century old, of Jewish assimilation into the allegedly superior German nation and into German cultural and economic life. As for Germany's position in international affairs, Hitler had long spoken of Germany's need for living space (*Lebensraum*) in the east. First, however, there was the continued need to break the chains of the hated Treaty of Versailles.

Whether the Nazis would ever get a chance to implement their ideological objectives depended, when Hitler became chancellor, upon whether they would be able to tighten their initially tenuous hold on the reins of power. Liberals, socialists, and communists remained bitterly opposed to Hitler; important segments of business, the army, and the churches were to varying degrees suspicious of the measures he might take. It was a combination, finally, of Hitler's daring and brutality, of the weaknesses of his opponents, and of numerous instances of extraordinary good luck that allowed him to establish his totalitarian dictatorship. When the Centre Party refused to join the Nazi-DNVP coalition in January 1933, Hitler demanded elections for a new Reichstag. The elections of March 5, 1933, were preceded by a brutal and violent campaign in which Nazi storm troopers under the command of Ernst Röhm figured prominently. Hitler was also able to take advantage of the Reichstag fire (probably the work of a lone and deranged Dutch communist) of February 27 to suspend civil liberties and arrest Communist as well as other opposition leaders. Despite this campaign of terror the Nazis did not win a majority, gaining only 43.9 percent of the total. The 8 percent acquired by the DNVP, however, was sufficient for the two parties to wield a majority in the Reichstag. At its first meeting on March 23 this Reichstag, under great pressure from the SA and Heinrich Himmler's SS (Schutzstaffel; "Protective Echelon"), voted in favour of an Enabling Act that gave Hitler power to ignore the constitution and to give his decrees the power of law.

The decree powers were the pseudolegal base from which Hitler carried out the first steps of the National Socialist revolution. Within two weeks of the passing of the Enabling Act, Nazi governors were sent out to bring the federal states into line, and a few months later the states themselves were abolished. On April 7, 1933, the civil service, including the universities, began to be purged of socialists, democrats, and Jews. On May 2 the trade unions were disbanded and replaced by what the Nazis called a Labour Front. In the meantime Göring had begun refashioning the political arm of the Prussian police into a secret political police (Gestapo [Geheime Staatspolizei]) to serve the Nazi cause, a process that was being duplicated by Himmler with the Bavarian police.

The brutality with which Hitler proved willing to meet any presumed challenge to his authority became dramatically evident in his ordering the June 30, 1934, murders of the SA leadership. Röhm's storm trooper street thugs had provided useful muscle during the party's long years of struggle, but their continuing penchant for unruliness, Hitler feared, could invite the army's intervention and therewith his own overthrow. To head off this possibility, Hitler engaged the loyal Himmler, who used his SS during the so-called Night of Long Knives to purge the SA of dozens of its top leaders, including Hitler's longtime friend Ernst Röhm. The penultimate step in Hitler's seizure of power came on Aug. 2, 1934, when, upon the death of President Hindenburg, he appropriated for himself the powers of the presidency and combined them with his own as chancellor. The final step came in February 1938 when Hitler took personal command over the three branches of the German armed forces.

The totalitarian state. The main purpose of the Nazi revolution, the goal to which power was to be put, was the establishment of the *Volksgemeinschaft*. Its creation required the purification and increase of the German race as well as its biological separation from the Jews, whose infusion of evil into the German bloodstream, the Nazis said, served to pollute and undermine Germany's well-being. Nazi efforts to purify the German race found a degree of scientific respectability from the new science of eugenics, the racial hygiene movement that flourished widely in the early decades of the 20th century. Nazi leaders spoke of their efforts to "reconstruct the German race." A Law for the Protection of Hereditary Health (July 14, 1933) allowed for the eventual sterilization of as many as two million people deemed unworthy of propagating. A Marriage Subsidy Law of July 1933 was calculated to stimulate the birth rate by granting loans to newly married couples; these loans would be forgiven incrementally with the birth of each additional child. The Nazi idealization of mothers and the celebration of motherhood as a special service to the Reich had the same objective. Hitler spoke of an eventual doubling of the German population through these measures. The most notorious of the steps taken to purify the German race was also a milestone

in the anti-Jewish legislation promulgated by the Nazis: the infamous Nürnberg Laws of September 1935, which forbade marriage or sexual relations between Jews and Germans and assigned to Jews a lower class of citizenship.

Nazi efforts to solve the "Jewish problem" were ultimately products of a vicious anti-Semitism that propelled the Nazi system toward increasingly extreme measures of persecution. SA terrorism, legislation expelling Jews from the civil service and universities, boycotts of Jewish businesses and professional people, and the eventual expropriation of Jewish-owned properties had by 1938 led to the emigration of roughly half of the 1933 Jewish population of 500,000 people. Until the massive display of violence against Jews and Jewish property known as Kristallnacht (Night of Broken Glass), during the night of Nov. 9–10, 1938, it still seemed possible for some Jews to remain in Germany, albeit in severely circumscribed circumstances. With Kristallnacht the Nazis ushered in a new level of persecution. In its aftermath Hitler put Hermann Göring in charge of Jewish policy. It became Göring's job to coordinate the numerous party and governmental agencies competing for control over—and profit from—the persecutions of the Jews.

Kristallnacht

To exclude Jews from the *Volksgemeinschaft* was as critical a need as attracting the German working classes to it, thereby undoing the long-standing alienation of the largely socialist-minded German worker from the nationalist consensus. Any success in this effort depended heavily upon the Nazis' ability to provide employment for the millions of jobless whom they inherited in 1933. Public spending on rearmament and on public works projects such as the superhighway, or autobahn, network helped create many of the jobs so desperately needed. By 1937 Germany was beginning to suffer from a labour shortage. It was equally important to inculcate workers with a sense of being an integral part of a racially based national community. For this the Nazis devised an elaborate program of subsidies for leisure-time activities for workers. Called "Strength through Joy," the program subsidized workers' vacations; it made possible excursions to mountain or seaside resorts and offered the possibility of cruises in the Mediterranean or the Baltic Sea. By providing leisure activities for the workers until then reserved for their economic and social superiors, the government attempted to integrate them into the *Volksgemeinshaft*. Workers were also given the opportunity to purchase an automobile, the Volkswagen, a symbol of wealth and status still largely reserved in the early 1930s for the upper classes. Hitler once said that Henry Ford had done more than anyone else to obliterate class differences in America.

Foreign policy. Hitler kept tight control over foreign affairs, formulating himself both the strategy and the tactics calculated to achieve his goals. The immediate objective was to reestablish Germany's position in world affairs; by this Hitler meant ending the humiliations attending the Treaty of Versailles, such as the demilitarized Rhineland and the limitations on German armaments. The chains of that treaty needed to fall with a loud clang, he said. The larger objective, the one he had spoken about since his entry into politics in the early 1920s, was the conquest for Germany of *Lebensraum*. Hitler believed that this space needed to be acquired in the east, at the expense of the Soviets, so as to secure for Germany the Ukrainian "breadbasket" and open up vast territories for German colonization. Hitler found justification for such conquests in his notions of German racial superiority over the Slavic peoples who inhabited the lands he coveted. Furthermore, he saw the Bolsheviks who now controlled Russia as the vanguard of the world Jewish conspiracy. Control of this territory was to become the foundation for Germany's economic and military domination of Europe and eventually, perhaps, of the world.

No such domination or expansion was possible without war, of course, and Hitler did not shrink from its implications. His rearming of Germany, begun in secret in 1933, was made public in March 1935 when he announced the creation of an air force and the reintroduction of general military conscription to provide the manpower for 36 new divisions in the army. In June of that same year he signed an agreement with the British that allowed a German naval buildup of up to 35 percent of Britain's surface naval strength and up to 45 percent of its tonnage in submarines. On March 7, 1936, he moved German forces into the demilitarized Rhineland. Versailles was dead. Neither the British nor the French had lifted a finger in its defense, choosing instead to sign agreements expressly negating the terms of Versailles. By 1936 Hitler was spending 10.2 billion marks on rearmament, and Göring was placed in charge of a so-called Four-Year Plan to prepare the German economy for war. On Nov. 5, 1937, Hitler gathered his general staff and admonished them to be prepared for war in the east no later than 1942 or 1943.

As a cover for his true intentions during the first years of power, Hitler spoke long and often of his desire for peace. All he wanted, he said, was a Germany allowed its rightful place in world affairs. As evidence of his pacific intentions he signed in January 1934 a 10-year nonaggression pact with Poland. A truer picture of his intentions became evident in July 1934, however, when Hitler encouraged the Nazi party in Austria to attempt an overthrow of the government of Chancellor Engelbert Dollfuss. Dollfuss was shot and killed, but the coup attempt was badly managed. Benito Mussolini's movement of Italian troops to the Austrian border forced Hitler to back away from supporting his Austrian partners.

It was from behind a cloak of anti-Bolshevism, carefully maintained by Propaganda Minister Joseph Goebbels, that Hitler managed to delude France and Britain, as well as the United States, leery of being drawn into another European quarrel, into accepting his claim to be the West's last bulwark against Bolshevik expansion. His agreement with the Japanese in 1936, the Anti-Comintern Pact, was directed against the Third Communist International. A year later Mussolini, after a year of German-Italian cooperation in aiding General Francisco Franco's rebel forces in the Spanish Civil War, added his signature to the pact.

Hitler made two dramatic foreign policy moves in 1938 that helped clarify to the world the extent of his less-than-pacific intentions. In March he annexed Austria to the Reich, justifying the move as a fulfillment of the principle of German national self-determination. Britain and France stood by quietly at this additional repudiation of the Versailles treaty. Next, Hitler engineered a diplomatic crisis with Czechoslovakia, claiming Czech mistreatment of its German minority in the Sudetenland. Against considerable opposition from his own military, Hitler was determined to go to war with the Czechs. Only through the intervention of Britain's Prime Minister Neville Chamberlain, who offered to come to Germany to appease Hitler and who managed to persuade the Czechs to yield to all of Hitler's demands, was war avoided. Chamberlain's intervention resulted in a four-power conference at Munich in late September 1938, at which the Italians, the French, and the British ceremoniously handed the Sudetenland over to Hitler. In diplomatic parlance ever since, the word "Munich" has come to symbolize caving in to the demands of a dictator.

Annexation of Austria

Certain that Britain and France would do nothing to stop him, Hitler, after Munich, decided to move up his timetable for conquests in the east. On March 15, 1939, Hitler seized what remained of Czechoslovakia, reshaping its pieces into a Bohemian and Moravian protectorate and a nominally independent state of Slovakia. Within a week he annexed from Lithuania the city of Memel and the surrounding countryside, territory lost to Germany as a consequence of Versailles. When Britain and France countered these moves by issuing their guarantee to Poland, clearly next on Hitler's agenda, a furious Hitler ordered his military to prepare an invasion of Poland. Poland was critical to Hitler's long-range strategy for the conquest of *Lebensraum* in the east; any invasion of the Soviet Union required that Polish territory be available as a staging area. Until the British-French guarantee he had hoped to enlist Poland, mostly through bombast and threat, as an ally in an attack on the Soviet Union. After the defeat of the Soviet Union, he believed Poland could be dealt with summarily. When Poland refused to play the role he had assigned it, Hitler began looking for allies in his resolution

of the Polish problem. In late May 1939 he signed his Pact of Steel with Mussolini. Then came the most sensational diplomatic rapprochement of the 20th century. Shortly after signing the pact with Italy, Hitler put out feelers to Stalin about the possibility of dividing eastern Europe, Poland included, into German and Soviet spheres of influence. These negotiations led to the notorious German-Soviet Nonaggression Pact of Aug. 24, 1939, a pact that secretly sealed Poland's doom and delayed Hitler's unbending determination to conquer *Lebensraum* from the Soviets for a few years. Thus prepared, on Sept. 1, 1939, Hitler launched his invasion of Poland. Two days later Britain and France declared war on Germany.

World War II. World War II is appropriately called "Hitler's war." The first two years were so extraordinarily successful that Hitler came close to realizing his aim of establishing a German hegemony in Europe. But his victories did not prove to be the essential components of a strategic conception that secured victory in the long run. Nonetheless, the early successes were spectacular. After the defeat of Poland within a month, Hitler turned his attention westward. He believed that it was necessary to defeat Britain and France before he could again turn eastward to the territories that were to become the "living space" for his new empire. The attack on the western front began in the spring of 1940. Hitler took Denmark and Norway during the course of a few days in April, and on May 10 he attacked France, along with Luxembourg, Belgium, and The Netherlands. Once again his armies achieved lightning victories; Luxembourg, Belgium, and The Netherlands were overrun in a few days, and France capitulated on June 21. Only the British, now alone, obstructed Hitler's path to total victory in the west.

Britain, Hitler determined, could be taken out of the war with air power. German bombers began their attack in August 1940, but the British proved intractable. The vaunted German air force failed to bring Britain to its knees partly because of the strength of the British air force, partly because the German air force was ill-equipped for the task, and partly because the British had been able since February 1940 to read German code. Yet Hitler had been so confident of a quick victory that even before the attack began he had ordered his military planners to draw up plans for an invasion of the Soviet Union. The date he set for the invasion was May 15, 1941.

Although the defeat of the Soviet Union was central to Hitler's strategic objective, he twice during the early months of 1941 allowed himself to be sidetracked into conflicts that delayed his invasion. In both instances he felt obliged to rescue his ally Mussolini from military difficulties. Mussolini had invaded Greece in October 1940, despite the fact that he was already in difficulty in North Africa, where he was unable to cut off Britain's Mediterranean lifeline in Egypt. In February 1941 Hitler decided to reinforce Mussolini in North Africa by sending an armoured division under the command of General Erwin Rommel. When Mussolini's invasion of Greece also bogged down, Hitler again decided to send reinforcements. To reach Greece, German troops had to be sent through the Balkan countries, all of them officially neutral. Hitler managed to bully these countries into accepting the passage of German troops, but on March 27 a coup in Yugoslavia overthrew the government, and the new rulers reneged on the agreement. In retaliation Hitler launched what he called Operation Punishment against the Yugoslavs. Yugoslav resistance collapsed quickly, but the effect was to delay for another month the planned invasion of the Soviet Union.

Invasion of the Soviet Union

When the invasion of the Soviet Union finally came, on June 22, 1941, it did so with both campaigns against the British, across the English Channel and in the Mediterranean, still incomplete. Hitler was prepared to take the risk that fighting on multiple fronts entailed because he was convinced that the war against the Soviet Union would be over by the onset of the Russian winter. The spectacular German advances during the first weeks of the invasion seemed proof of Hitler's calculation. On July 3 his army chief of staff wrote in his war diary that the war had been won. The German Army Group North was approaching Leningrad; Army Group Centre had broken through the Soviet defenses and was rushing toward Moscow; and Army Group South had already captured vast reaches of the Ukraine. The prospect of capturing the summer harvest of the Ukraine along with the oil fields of the Caucasus led Hitler to transfer troops driving toward Moscow to reinforce those operating in the south.

Hitler's generals later considered this decision a turning point in the war. The effect was to delay until October the drive toward Moscow. By then an early winter had set in, greatly impeding the German advance and finally bringing it to a halt at the outskirts of Moscow in early December. Then, on December 6, the Soviets, having had time to regroup, launched a massive counteroffensive to relieve their capital city. On the following day the Japanese, nominally Germany's ally, launched their attack on the U.S. naval base at Pearl Harbor in Hawaii. Although they had not bothered to inform Hitler of their intentions, he was jubilant when he heard the news. "Now it is impossible for us to lose the war," he told his aides. On December 11 he declared war on the United States.

Though his plans for a quick defeat of the Soviet Union had not been realized, Hitler's troops at the end of 1941 controlled virtually all of the European territory of the Soviet Union. They stood at the outskirts of Leningrad and Moscow and were in control of the entire Ukraine. To prepare for what would now have to be the campaign of 1942, Hitler dismissed a number of generals and assumed himself the strategic and operational command of the armies on the eastern front.

At the high point of Hitler's military successes in the Soviet Union, members of the Nazi leadership were, with Hitler's understanding, feverishly planning for the new order they intended to impose on the conquered territories. Its realization called both for the removal of obstacles to German settlements and for a solution to the "Jewish problem." Nazi planners were drafting an elaborate scheme, General Plan East, for the future reorganization of eastern Europe and the western Soviet Union which called for the elimination of 30 million or more Slavs and the settlement of their territories by German overlords who would control and eventually repopulate the area with Germans. During the fall of 1941 Heinrich Himmler's SS expanded and refurbished with gas chambers and crematoriums an old Austrian army barracks near the Polish rail junction at Auschwitz. Here was to continue, with greater efficiency, the mass murder of Jews that had begun with the June invasion, when SS murder squads began rounding up Jews and shooting them by the thousands. The assurance of victory in the east, the heartland of European Jewry, gave Nazis the confidence that the "Jewish problem" could be rendered amenable to a "final solution." Experts suggest that ultimately somewhere between five and six million Jews were murdered in the death factories of eastern Europe. At least an equal number of non-Jews died of murder and starvation in places like Auschwitz, including two and a half million Soviet prisoners of war and countless others from the eastern European nationalities.

The success of Nazi armies until the end of 1941 had made it possible to spare German civilians on the home front from the misery and sacrifices demanded of them during World War I. Hitler's imagination, however, was haunted by the memory of the collapse of the home front in 1918, and to avoid a repetition the Nazis undertook to loot the occupied territories of food and raw materials as well as labour. Food shortages in Germany were not serious until late in the war. Women were allowed to stay at home, and the energies of the German work force were not stretched to their limits because eventually some seven million foreign slave labourers were used to keep the war effort going.

Through much of 1942 an ultimate German victory still seemed possible. The renewed offensive in the Soviet Union in the spring at first continued the successes of the previous year. Once again Hitler chose to concentrate on the capture of the Caucasus and its oil at the expense of the Moscow front. The decision made inevitable a major battle over the industrial centre at Stalingrad (now Volgograd). Elsewhere, by midsummer of 1942, Rommel's

Afrika Korps advanced to within 65 miles of Alexandria in Egypt. In the naval battle for control of the Atlantic sea lanes, German submarines were able to maintain their ability to intercept Allied shipping into mid-1943.

By early 1943, however, the tide had clearly begun to turn. The great winter battle at Stalingrad brought Hitler his first major defeat. His entire Sixth Army was killed or captured. In North Africa Rommel's long success ended in late 1942 when the British broke through at El Alamein. At the same time a joint British-American force landed in northwestern Africa on the coast of Morocco and Algeria. By May 1943 the German and Italian forces in North Africa were ready to surrender. That same summer the Allies broke the back of the German submarine campaign in the Atlantic. On July 10 the Allies landed in Sicily. Two weeks later Mussolini was overthrown, and in early September the Italians withdrew from the war.

The addition of an Italian front made the rollback of German forces on all fronts that much more likely. In the Soviet Union, German forces were stretched across 2,500 miles. They had lost their air superiority when Allied bombing raids on German cities forced the withdrawal of large numbers of fighter planes. Allied bombings reached a high point in midsummer when a raid on Hamburg killed 40,000 of its inhabitants. Other cities suffered similar raids. Shortages of food, clothing, and housing began to afflict German cities as inevitably as did the Allied bombers.

The rollback of German forces continued inexorably during 1944. On June 6 the Allies in the west launched their invasion of France across the English Channel. In the east the Soviet Army was advancing along the entire 2,500-mile front. By the end of the year it stood poised on the eastern frontiers of prewar Germany. In the west, British and American troops stood ready to attack across the western borders.

On the German home front, 1944 became a year of acute suffering. On July 20 an officers' plot, part of a long-simmering opposition to Hitler from within German military and civilian circles, was carried out, but Hitler managed to escape the dramatic attempt on his life practically unharmed. He attributed his survival to his having been selected by fate to succeed in his mission of restoring Germany to greatness.

Fate did not again intervene on Hitler's behalf. In mid-January of 1945 he withdrew underground into his bunker in Berlin where he remained until his suicide on April 30. By that time Soviet soldiers were streaming into Berlin. All that remained of the Reich was a narrow wedge of territory running southward from Berlin into Austria. (For further discussion of World War II, see WORLD WARS, THE.)

With the Soviet army in control of Berlin and the western Allies within striking distance to the west and the south, there was no prospect of dividing them. Nonetheless, when Hitler's successor, Grand Admiral Karl Dönitz, sought to open negotiations for a surrender a few days after Hitler's death, he still hoped that a separate surrender to the British and Americans in the west might allow the Reich to rescue something from the Soviets in the east. The western Allies, fearful of any move that might feed the suspicions of Stalin, refused to consider the German proposal, insisting that a German surrender be signed with all of the Allies at the same time. Early in the morning of May 7, 1945, a German delegation came to U.S. General Dwight D. Eisenhower's headquarters in Rheims, Fr., and at 2:41 AM signed the surrender documents. Despite the fact that a Soviet major general signed for the Soviet Union, Stalin insisted that a second surrender ceremony be arranged in Soviet-occupied Berlin. This second surrender was signed in a Berlin suburb the following afternoon. (K.A.Sc.)

The era of partition

As a legacy of unconditional surrender at the end of World War II, a truncated Germany was divided into four zones of Allied military occupation; Berlin, the capital between 1871 and 1945 of the unified German Reich, was similarly divided. In 1949 the zones of occupation of France, the United Kingdom, and the United States were merged to form the Federal Republic of Germany (Bundesrepublik Deutschland; commonly known as West Germany), with its capital at Bonn. The three western sectors of Berlin also had a separate administration, beginning in 1948, constituting West Berlin. West Berlin's political affiliation was with West Germany, although it was a detached exclave of territory at a distance of 110 miles (180 kilometres). Also in 1949 the Soviet zone became the German Democratic Republic (Deutsche Demokratische Republik; commonly known as East Germany), with its capital in the former Soviet sector of Berlin (*i.e.*, East Berlin).

West German recovery

West Germany's recovery from total economic and political prostration and the devastation of its cities and capital industries at the end of World War II was of such dramatic proportions as to become a modern legend. The *Wirtschaftswunder* ("economic miracle") of the 1950s and '60s made it the world's fourth largest economy after the United States, the Soviet Union, and Japan and one of the world's largest trading nations.



Germany, 1952–90.

Domestically, the West German state grew into one of the most stable of Western democracies, with three major political parties competing peacefully for power. In large measure because of the 5 percent of votes required for representation in the federal parliament (Bundestag) and provincial diets, extremist parties of the right and left did not succeed in West German politics. The West Germans as citizens moved from their accustomed political apathy to becoming intensely involved and well-informed on the issues that affected their lives.

East Germany was established as a socialist state of the Stalinist type and became a member of the Council for Mutual Economic Assistance (Comecon) and of the Warsaw Pact military alliance. It had the most advanced economy and the highest living standards of all the Soviet-bloc states, but industrially it lagged behind the Western level. To outside observers, the centralized and dictatorial regime of the Socialist Unity Party (a forced union of the Social Democratic Party with the Communists), buttressed by the omnipresent State Security Police (Stasi), appeared unassailable. But the precarious nature of this foundation was demonstrated by the East German government's swift and unforeseen collapse under popular pressure in late 1989, most strikingly symbolized by the opening of the Berlin Wall on November 9. At free elections on March 18, 1990, the majority of East Germans, by voting for the counterpart of the West German Christian Democratic Union, signaled their wish to join the Federal Republic at the earliest possible moment (see below, *The reunification of Germany*). (T.H.El./G.H.K./Ed.)

ALLIED OCCUPATION AND THE FORMATION OF THE TWO GERMANIES, 1945–49

Following the German military leaders' unconditional surrender in May 1945, the country lay prostrate. The German state ceased to exist, and sovereign authority passed to the victorious Allied powers. The physical devastation from Allied bombing campaigns and from ground battles was enormous: an estimated one-quarter of the nation's housing was destroyed or damaged beyond use, and in many cities the toll exceeded 50 percent. Germany's economic infrastructure had largely collapsed as factories and transportation systems ceased to function. Rampant inflation was undermining the value of the currency, and an acute shortage of food reduced the diet of many city dwellers to the level of malnutrition. These difficulties were compounded by the presence of millions of homeless German refugees from the east. The end of the war came to be remembered as "zero hour," a low point from which virtually everything had to be rebuilt anew from the ground up.

For purposes of occupation, the Americans, British, French, and Soviets divided Germany into four zones. The American, British, and French zones together made up the western two-thirds of Germany, while the Soviet zone comprised the eastern third. Berlin, the former capital, was placed under joint four-power authority but was partitioned into four sectors for administrative purposes. An Allied Control Council was to exercise overall joint authority.

Allied zones of occupation

Not all of prewar Germany was included in these arrangements. The Soviets unilaterally severed the German territories east of the Oder and Neisse rivers and placed these under the direct administrative authority of the Soviet Union and Poland, with the larger share going to the Poles as compensation for territory they lost to the Soviet Union. The former provinces of East Prussia, Pomerania, and Silesia were thus stripped from Germany. Since virtually the entire German population of some 9.5 million in these and adjacent regions was expelled westward, this amounted to a de facto annexation of one-quarter of Germany's territory as of 1937, the year prior to the beginning of German expansion under Hitler. The Western Allies acquiesced in these actions by the Soviets, taking consolation in the expectation that their postwar territorial dispositions were merely temporary expedients that would quickly be superseded by the final peace terms.

As a result of irreconcilable differences among the Allied powers, however, no peace conference was ever held. The issue of German reparations proved particularly divisive. The Soviet Union, whose population and territory had suffered terribly at the hands of the Germans, demanded large-scale material compensation. The Western Allies initially agreed to extract reparations but soon came to resent the Soviets' seizures of entire German factories as well as current production. Under the terms of inter-Allied agreements, the Soviet zone of occupation, which encompassed much of German agriculture and was less densely populated than those of the other Allies, was to supply foodstuffs to the rest of Germany in return for a share of reparations from the western occupation zones. But when the Soviets failed to deliver the requisite food, the Western Allies found themselves forced to feed the German population in their zones at the expense of their own taxpayers. The Americans and British therefore came to favour a revival of German industry so as to enable the Germans to feed themselves, a step the Soviets opposed. When the Western powers refused in 1946 to permit the Soviets to claim further reparations from their zones, cooperation among the wartime allies deteriorated sharply.

During the postwar years two quite different economic and social systems took shape. In the western zones the Allies permitted a market economy based on private property to reemerge, though at the outset under extensive regulation. The emergence of a free and pluralistic press was also fostered. In carrying out the policy of de-Nazification agreed upon by all four Allied powers, the Western occupation authorities sought to punish those Germans who had been deeply involved in the Nazi regime while levying lighter penalties on those less implicated. The Soviets proceeded against former Nazis in their occupation zone more summarily and with little attempt at differentiating among degrees of complicity. They subjected the press and all other means of communication in their zone to increasingly close censorship. In the economic sphere they nationalized most industries without compensation for the previous owners, and they effected a sweeping agrarian reform by expropriating large estates and distributing the land among small farmers.

Beginning in the summer of 1945, the occupation authorities permitted the formation of German political parties preparatory to elections for new local and regional representative assemblies. Two of the major leftist parties of the Weimar era quickly revived: the moderate Social Democratic Party (SPD) and the German Communist Party (KPD), which was loyal to the Soviet Union. These were soon joined by a new creation, the Christian Democratic Union (CDU), with its Bavarian sister party, the Christian Social Union (CSU). The leaders of this Christian Democratic coalition had for the most part been active in the moderate parties of the Weimar Republic, especially the Catholic Centre Party. They sought to win popular support on the basis of a nondenominational commitment to Christian ethics and democratic institutions. Liberal-minded Germans who favoured a secular state and laissez-faire economic policies formed a new Free Democratic Party in the western zones and a Liberal Democratic Party in the Soviet zone. Numerous smaller parties were also launched in the western zones.

Postwar political parties

In April 1946, under pressure from the occupation authorities, the Social Democratic leaders in the Soviet zone agreed to merge with the Communists, a step denounced by the Social Democrats in the western zones. The resulting Socialist Unity Party (SED) swept to victory with the ill-concealed aid of the Soviets in the first elections for local and regional assemblies in the Soviet zone. However, when in October 1946 elections were held under fairer conditions in Berlin, which was under four-power occupation, the SED tallied fewer than half as many votes as the Social Democratic Party, which had managed to preserve its independence in the old capital. Thereafter, the SED, which increasingly fell under communist domination as Social Democrats were systematically purged from its leadership ranks, avoided free, competitive elections by forcing all other parties to join a permanent coalition under its leadership.

The occupying powers soon approved the formation of regional organs of self-administration called *Länder* (singular *Land*), or "states." By 1947 those *Länder* established in the western zones had freely elected parliamentary assemblies. Institutional developments followed a superficially similar pattern in the Soviet zone, but there the political process remained less than free because of the dominance of the Soviet-backed SED.

When it had become apparent by 1947 that the Soviet Union would not permit free, multiparty elections throughout the whole of Germany, the Americans and British amalgamated the German administrative organs in their occupation zones in order to foster economic recovery. The resulting unit, called Bizonia, operated through a set of German institutions located in the city of Frankfurt am Main. Its federative structure would later serve as the model for the West German state.

Creation of West Germany

In the politics of Bizonia, the Social Democrats and the Christian Democrats quickly established themselves as the major political parties. The Social Democrats held to their long-standing commitment to nationalization of basic industries and extensive government control over other aspects of the economy. The Christian Democrats, after initially inclining to a vaguely conceived "Christian socialism," swung to espousal of a basically free-enterprise orientation. In March 1948 they joined with the laissez-faire Free Democrats to install as architect of Bizonia's economy Ludwig Erhard, a previously obscure economist who advocated freeing the economy from government control.

When repeated meetings with the Soviets failed to produce four-power cooperation, the Western occupying powers decided in the spring of 1948 to move on their own.

They were particularly concerned about the deteriorating economic conditions throughout occupied Germany, which burdened their own countries and awakened fears of renewed political extremism among the Germans. The Western powers therefore decided to extend to their occupation zones American economic aid, which had been instituted elsewhere in western Europe a year earlier under the Marshall Plan. To enhance the effectiveness of that aid, the Americans, British, and French effected a currency reform in their zones that replaced Germany's badly inflated currency with a new, hard *Deutsche Mark,* or DM. Western Germany's economy responded quickly, as goods previously unavailable for nearly worthless money came onto the market.

The Soviets responded angrily to the currency reform, which was undertaken without their approval. When the new *Deutsche Mark* was introduced into Berlin, they protested vigorously and boycotted the Allied Control Council. Then, in June 1948 they blockaded the western sectors of the old capital, which were surrounded by territory occupied by the Soviet Red Army and lay about 100 miles from the nearest Western-occupied area. By sealing off the railways, highways, and canals used to deliver food and fuel, as well as the raw materials needed for the factories of a city of more than two million people, the Soviets sought to paralyze West Berlin and drive out their erstwhile allies. They were thwarted, however, when the Western powers mounted an around-the-clock airlift that supplied the West Berliners with food and fuel throughout the bleak winter of 1948–49. In May 1949 the Soviets relented and lifted the blockade.

Berlin blockade

Formation of the Federal Republic of Germany. Instead of halting progress toward the political integration of the western zones, as the Soviets apparently intended, the Berlin blockade accelerated it. In April 1949 the French began to merge their zone into Bizonia, which became Trizonia. That September a Parliamentary Council of 65 members chosen by the parliaments of the *Länder* began drafting a constitution for a West German government. Twenty-seven seats each in this council were held by the Social Democrats and the Christian Democrats, five by the Free Democrats, and the rest by smaller parties, including two by the Communists. The Council completed its work in the spring of 1949, and the Federal Republic of Germany came into being in May 1949 after all the *Länder* except Bavaria had ratified the Basic Law (*Grundgesetz*), as the constitution was designated to underline the provisional nature of the new state. This document specified that it was designed only for temporary use until a constitution had been freely adopted by the German people as a whole.

The Basic Law was approved by the Western Allied military governors with certain reservations, notably the exclusion of West Berlin, which had been proposed as the 12th *Land* of the federation. The 11 constituent *Länder* of West Germany, then, were Bavaria, Bremen, Hamburg, Hesse, Lower Saxony, North Rhine-Westphalia, Rhineland-Palatinate, Schleswig-Holstein, Baden, Württemberg-Baden, and Württemberg-Hohenzollern (the last three were merged in 1952 to form Baden-Württemberg, and in 1957 Saarland became the 10th *Land*).

By the terms of the Basic Law, the Federal Republic of Germany was established with its provisional capital in the small university city of Bonn. The West German state took shape as a federal form of parliamentary democracy. An extensive bill of rights guaranteed the civil and political freedoms of the citizenry. In keeping with German traditions, many spheres of governmental authority were reserved for the individual *Länder*. The key locus of power on the federal level lay in the lower legislative chamber, the Bundestag, elections for which had to take place at least every four years. The deputies were chosen by a voting procedure known as "personalized proportionality" that combined proportional representation with single-seat constituencies by according each voter two ballots—one for a party and one for a specific candidate. The composition of the chamber was ultimately determined by the percentage of the national vote each party received. But half the seats were filled by candidates who commanded the most votes in the constituencies, thereby giving the citizenry local advocates in the legislature. In order to minimize the proliferation of smaller political parties that had helped to discredit democracy in the Weimar Republic, a party had to win a minimum of 5 percent of the overall vote to gain representation in the Bundestag. All citizens 18 years of age and older were eligible to vote (until 1971 the minimum age was 21).

West German governmental structure

The Bundestag elected the head of government, the federal chancellor. Once installed, the chancellor chose members of the Cabinet, who presided over the ministries, and the chancellor determined and assumed responsibility for general policy. To rule out the instability that characterized the Cabinets of the Weimar Republic, the Basic Law introduced a novel provision designed to strengthen the constitutional position of the chancellor. It required that a majority vote of no-confidence in the Bundestag could unseat the chancellor and his Cabinet only if it was accompanied by a positive majority in favour of a substitute candidate for the chancellorship.

Whereas the chancellor headed the government, a federal president stood at the head of the state. Because of the misuse of presidential power in the Weimar Republic, the powers of this office were greatly reduced by the Basic Law. The president was not directly elected, as in Weimar, by popular vote, but instead was chosen indirectly by a federal convention. Half of that body, which came into being solely to select a president, was composed of members of the Bundestag, and the other half was chosen by the parliaments of the *Länder*. The president represented the Federal Republic in its formal dealings with other countries, accrediting and receiving envoys. In domestic affairs, presidential authority was limited to nominating candidates for the chancellorship and dissolving the Bundestag to make way for new elections in the event that no majority could be found to install a chancellor.

The *Länder* were represented in the upper legislative chamber, the Bundesrat. Its members were not elected but rather designated by the governments of the *Länder,* the number varying according to the states' populations. Constitutional changes required a two-thirds vote in both the Bundesrat and the Bundestag. On all legislation affecting the rights of the *Länder* the Bundesrat had an absolute veto. On other legislation a majority of the Bundesrat could block enactment of bills passed by the Bundestag by what is known as a "suspensive veto." If the margin in the Bundesrat was a simple majority, this could be overridden by a majority vote in the Bundestag, but if the margin was two thirds, a similar vote was necessary to override.

The final key institution of the Federal Republic was the Federal Constitutional Court. Independent of both the legislative and executive branches, it successfully introduced into German practice for the first time the American principle of judicial review of legislation. Half of its members were named by the Bundestag, the other half by the Bundesrat, each for a nonrenewable term of 12 years. Its seat was in the city of Karlsruhe.

Initially, the Federal Republic was not a sovereign state. Its powers were circumscribed by an Occupation Statute drawn up by the American, British, and French governments in 1949. That document reserved to those powers ultimate authority over such matters as foreign relations, foreign trade, the level of industrial production, and all questions relating to military security. Only with the permission of the Western occupation powers could the Federal Republic legislate or otherwise take action in those spheres. Alterations in the Basic Law required the unanimous consent of the three Western powers, and they reserved veto power over any legislation they deemed unconstitutional or at variance with occupation policies. In the event of an emergency that endangered the new West German government, the Western Allies retained the right to resume their full authority as occupying powers.

Formation of the German Democratic Republic. When it became clear that a West German government would be established, a so-called election for a People's Congress was held in the Soviet occupation zone in May 1949. But instead of choosing among candidates, voters were allowed only the choice of approving or rejecting—usually

in less-than-secret circumstances—"unity lists" of candidates drawn from all parties, as well as representatives of mass organizations controlled by the communist-dominated SED. Two additional parties, a Democratic Farmers' Party and a National Democratic Party, designed to attract support, respectively, from farmers and from former Nazis, were added with the blessing of the SED. By ensuring that communists predominated in these unity lists, the SED determined in advance the composition of the new People's Congress. According to the official results, about two-thirds of the voters approved the unity lists. In subsequent elections, favourable margins in excess of 99 percent were routinely announced.

In October 1949, following the formation of the Federal Republic, a constitution ratified by the People's Congress went into effect in the Soviet zone, which became the German Democratic Republic (East Germany), with its capital in the Soviet sector of Berlin. The People's Congress was renamed the People's Chamber, and this body, together with a second chamber composed of officials of the five *Länder* of the Soviet zone (which were abolished in 1952 in favour of centralized authority), designated the communist Wilhelm Pieck of the SED as president of the German Democratic Republic on Oct. 11, 1949. The next day, the People's Chamber installed the former Social Democrat Otto Grotewohl as premier at the head of a Cabinet that was nominally responsible to the chamber. Although the German Democratic Republic was constitutionally a parliamentary democracy, decisive power actually lay with the SED and its boss, the veteran communist functionary Walter Ulbricht, who held only the obscure position of deputy premier in the government. In East Germany, as in the Soviet Union, the government served merely as the agent of an all-powerful communist-controlled party, which was in turn ruled from above by a self-selecting Politburo.

The power of the party

POLITICAL CONSOLIDATION AND ECONOMIC GROWTH, 1949–69

The government that emerged from the Federal Republic's first general election in August 1949 was based on a coalition of the Christian Democrats with the Free Democrats. Konrad Adenauer of the Christian Democratic Union, a veteran Roman Catholic politician from the Rhineland, was elected the country's first chancellor by a narrow margin in the Bundestag. Because of his advanced age of 73, Adenauer was expected by many to serve only as an interim officeholder, but in fact he retained the chancellorship for 14 years. Elected as West Germany's first president was Theodor Heuss of the Free Democratic Party. As economics minister in the Adenauer Cabinet, Ludwig Erhard launched the Federal Republic on a phenomenally successful course of economic revival under the formula of "social market economics," or welfare-state capitalism. His policies left the means of production in private hands and allowed market mechanisms to set price and wage levels. Social justice was served by means of government measures designed to ensure an equitable distribution of the wealth generated by the pursuit of profit. So successful were these modified free-market policies that West Germany soon abolished all rationing and became renowned for its "economic miracle" as industrial output rapidly recovered and living standards steadily rose.

One of the most urgent internal problems for Adenauer's first administration was the resettlement of refugees. By 1950 the Federal Republic had become the new home of 4.5 million Germans from the territory east of the Oder-Neisse line; 3.4 million ethnic Germans from Czechoslovakia, prewar Poland, and other eastern European countries; and 1.5 million from the German Democratic Republic. The presence of these refugees put a heavy social burden on the Federal Republic, but their assimilation proved surprisingly easy. Many of the refugees were skilled, enterprising, and adaptable, and their labour proved an important factor in West Germany's economic recovery.

East Germany's internal policies

In the German Democratic Republic, the SED regime accorded priority to building a viable economy in a territory that lacked rich natural resources, was less than one-half the size of the Federal Republic, and had a population (17 million) only one-third as large. The regime used its centralized control over a planned economy to invest heavily in the construction of basic industries at the expense of the production of consumer goods. Moreover, war reparations required that much productive capacity be diverted to Soviet needs. Despite an impressive rate of industrial growth, the standard of living remained low, lagging far behind that of West Germany. Even food was a problem, as thousands of farmers fled to West Germany each year rather than give in to mounting pressure to merge their land (which many had only recently obtained through the postwar agrarian reform) into the collective farms favoured by the communist regime. Food rationing had to be continued long after it had ended in West Germany. The resulting material hardships, along with relentless ideological indoctrination, repression of dissent, and harassment of churches by a militantly atheistic regime prompted many thousands of East Germans to flee to West Germany every year. In 1952 East Germany sealed its borders with West Germany, but East Germans continued to leave through Berlin, where free movement still prevailed between the four separate occupation sectors of the city.

In the Federal Republic, Adenauer followed a resolute policy of linking the new West German state closely with the Western democracies, even at the cost of perpetuating Germany's division for the time being. In 1951 the chancellor succeeded in gaining membership for West Germany in the European Coal and Steel Community, which later served as the core of the European Communities (EC). In that same year the Americans, British, and French agreed to a revision of the Occupation Statute that substantially increased the internal authority of the Federal Republic. Skillfully exploiting the Western fears of a communist assault on Europe that had been awakened by the Korean War, Adenauer gained further concessions from the Western occupying powers in return for his agreement to rearm West Germany within the context of a western European defense system. In 1955 the Federal Republic became a full member of the North Atlantic Treaty Organization (NATO) and gained sovereignty over its foreign relations as the Occupation Statute expired. In 1957 it became one of the charter members of the European Economic Community.

West German foreign relations

With West Germany's economic recovery continuing impressively, the voters confirmed the policies of the Adenauer government: in the Bundestag election of 1953, the coalition of the Christian and Free Democrats increased its previously thin majority. In the election of 1957, the chancellor's party achieved the first absolute majority ever recorded by a German party in a free general election. The Free Democrats, who had left the government in 1956 because of policy disputes, remained in opposition thereafter, along with the Social Democrats.

Mounting dissatisfaction with the SED regime in East Germany led to the first popular uprising in the postwar Soviet bloc when workers in East Berlin, the seat of government, went on strike on June 17, 1953, to protest against increased production quotas. When the regime failed to respond, the workers took to the streets and demanded a change in government. The rebellion quickly spread throughout East Germany and was quelled only when Soviet troops intervened, killing at least 21 people and wounding hundreds of others. A wave of retribution followed, as some 1,300 were sentenced to prison for taking part in the uprising, which the East German government portrayed as a plot on the part of West Germany and the United States.

In the wake of the uprising, Joseph Stalin's successors in the leadership of the Soviet Union upgraded the status of the German Democratic Republic. In 1954 reparations to the Soviet Union were halted, and Moscow proclaimed East Germany a sovereign state. In 1955 East Germany became a charter member of the Warsaw Pact, the Soviet bloc's military alliance. The SED leadership loosened ideological controls on artistic and intellectual activities somewhat, and the production of consumer goods was increased. Pressure on farmers to enter collective farms was also relaxed. Agricultural yields improved, and the

last food rationing was ended in 1958. Within a few years, however, the government resumed its repressive measures and again shifted its economic priorities to favour the collectivization of agriculture and investment in heavy industry at the expense of consumer goods. The flight of refugees through Berlin continued, with a high proportion of technicians, managers, and professionals among them.

The Berlin Wall

In 1961 the flow of refugees to West Germany through Berlin increased dramatically, bringing the total number of East Germans who had fled since the war to some three million. On Aug. 13, 1961, the East German government surprised the world by sealing off East Berlin from West Berlin, first by barbed wire and later by construction of a concrete wall through the middle of the city. No one was thenceforth permitted to go to the West through the tightly guarded crossing points without official permission, which was rarely granted. East Germans who sought to escape by climbing over the wall risked being shot by East German guards under orders to kill, if need be, to prevent the crime of "flight from the republic." By thus imprisoning the population, the SED regime stabilized the economy of East Germany, which eventually became the most prosperous of the Soviet bloc but which nevertheless continued to lag behind that of West Germany in both the quantity and quality of its consumer goods. Under party boss Ulbricht, the East German government also tightened the repressive policies of what had become a totalitarian communist dictatorship. Upon the death of President Pieck in 1960, Ulbricht had assumed the powers of the presidency as head of a newly created Council of State. In 1968 he imposed a new constitution on East Germany that sharply curtailed civil and political rights.

In the Federal Republic's Bundestag election of 1961 the Christian Democrats suffered losses for the first time. The Social Democratic Party, which had broadened its appeal by jettisoning the last remnants of its Marxist past and accepting the existing economic system in its Bad Godesberg program of 1959, scored impressive gains. Adenauer managed to retain the chancellorship by forming another coalition with the Free Democrats, but his position was weakened. He had tarnished his image in 1959 when he announced his candidacy for the presidency only to withdraw in favour of the election of a lacklustre party colleague, Heinrich Lübke, when he realized that under the Basic Law the president had little power. The elderly chancellor was further weakened when in 1962 his defense minister, Franz Josef Strauss of the Bavarian Christian Social Union, adopted high-handed methods in bringing about the arrest of journalists of the popular weekly news magazine *Der Spiegel* in connection with an alleged security leak. At the insistence of the Free Democrats, Adenauer relinquished the chancellorship in October 1963 to Erhard. Although the Erhard Cabinet held its own in the election of 1965, the architect of the "economic miracle" himself fell from power in November 1966 when the Free Democrats withdrew their support because of disagreements over how to respond to a recession. For the next three years the Federal Republic was governed by a grand coalition of the two largest parties, the Christian Democrats and the Social Democrats, with Christian Democrat Kurt Georg Kiesinger as chancellor and Social Democrat Willy Brandt as foreign minister.

OSTPOLITIK AND RECONCILIATION, 1969–89

The social-liberal coalition

When the Social Democrats scored impressive gains in the Bundestag election of 1969 and also captured the presidency for their candidate, Gustav Heinemann, West Germany underwent its first full-scale change of government. After 20 years of Christian Democratic domination, the Social Democrats captured the chancellorship for Brandt in coalition with the Free Democrats, whose leader Walter Scheel became foreign minister. This so-called social-liberal coalition carried through a number of domestic reforms, but its principal impact was on the Federal Republic's relations with East Germany and the communist-ruled countries of eastern Europe. While confirming West Germany's commitment to the Western alliance, the new government embarked upon a bold new "eastern policy," or *Ostpolitik*.

Previously, the Federal Republic had refused to recognize even the existence of the East German government. And by the terms of the so-called Hallstein doctrine (named for one of Adenauer's key foreign policy aides, Walter Hallstein) the Bonn authorities had refused to maintain diplomatic relations with all those countries (other than the Soviet Union) that recognized the German Democratic Republic. Now the Brandt-Scheel Cabinet reversed these policies by opening direct negotiations with East Germany in 1970 for the purpose of normalizing relations between the two German states.

Brandt's *Ostpolitik*

In 1970 the social-liberal Cabinet entered into treaties with the Soviet Union and Poland that required Bonn to recognize the Oder-Neisse line as Germany's eastern boundary. After the Soviet Union joined in 1971 with the Americans, British, and French in a Four Power Agreement which regularized Berlin's status and opened the way for an easing of the West Berliners' lot, the Brandt-Scheel Cabinet reached agreement in 1972 with East Germany on the Basic Treaty, designed to regularize the relations of the two German states. By its terms each side recognized, and agreed to respect, the other's authority and independence. Each foreswore any title to represent the other internationally, which meant West Germany's abandonment of its long-standing claim to be the sole legitimate spokesman of the German people. The two agreed to exchange "permanent missions," which meant that their relations stopped short of full diplomatic recognition.

The new *Ostpolitik* met with bitter resistance within West Germany from the Christian Democrats, who denounced it as a surrender on many points that should await settlement by a peace treaty, including the status of the eastern territories severed from Germany in 1945. The Christian Democrats were especially offended by the prospect of the Federal Republic's according legitimacy to a dictatorial German Democratic Republic that refused to allow free elections, maintained the Berlin Wall, and ordered its border guards to shoot fleeing citizens. The Christian Democrats therefore pledged not to ratify the Basic Treaty if they were returned to power in the Bundestag election of November 1972. The voters endorsed the Brandt government's *Ostpolitik*, however, by making the Social Democrats the largest party in the Bundestag (for the first time) and by strengthening their coalition partner, the Free Democrats, as well. The Basic Treaty was signed at the end of 1972, and the following year both German states gained admission to the United Nations.

The Basic Treaty

When West Germany's original overtures toward East Germany had met with resistance from Ulbricht, the path for negotiations had been cleared by a withdrawal of Soviet support that led to Ulbricht's replacement as East German leader in 1971 by another communist functionary, Erich Honecker. In his last years, Ulbricht had experimented with a decentralization of economic decision making, but under Honecker the German Democratic Republic reverted to Soviet-style centralized planning.

East Germany benefited greatly from the Basic Treaty with the Federal Republic. Once Bonn had accorded East Germany recognition, the Western democracies followed suit, so that the East German state at last enjoyed the international acceptance it had so long sought. Economically, too, the Basic Treaty proved a boon to East Germany. Spurred by West German credits, trade between the two German states increased, yielding valuable West German currency for East Germany. The latter derived further income from annual fees paid to it by West Germany for western travelers' use of the highways through East Germany to Berlin, and from ransoms paid by West Germany for the release of political prisoners held in East Germany. The larger number of West Germans allowed to visit East Germany also brought in hard currency. Each year the East German government reported impressive leaps in productivity which, after that regime's collapse, proved to be largely fictional. In actuality, the material gap between the two parts of Germany widened. The East German government neglected to maintain the country's infrastructure in order to concentrate its resources on industrial production for export purposes; the results of this became increasingly apparent as East Germany's roads,

railways, and buildings deteriorated. An acute housing shortage also persisted. Waiting periods of years were still required for the purchase of major consumer items such as automobiles, which continued to be crudely manufactured according to standards of the early postwar period, while those of West Germany ranked high in the world for quality and advanced design.

Growing East-West contacts

The benefits of international recognition were offset by the dangers posed to the dictatorial East German government by increased visits by West Germans to the east. In an effort to deal with the subversive effects of such contacts, the East German government repeatedly sought to reduce the influx of West German visitors by raising the fees it charged for visas. It classified some two million of its citizens as "bearers of secrets" and forbade them personal contact with Westerners. To stifle dissent at home, the government tightened its already repressive ideological controls on artists and intellectuals, imprisoning some and stripping others of their citizenship and banishing them to West Germany. To emphasize the distinctness of the German Democratic Republic, the amended constitution of 1974 minimized the use of the word "German" and stressed the socialist nature of the East German state and its irrevocable links with the Soviet Union.

In West Germany, Brandt resigned in 1974 after one of his trusted aides was unmasked as a spy for East Germany. Brandt's successor as chancellor was the Social Democrat Helmut Schmidt, who continued the coalition of the Social Democrats with the Free Democrats. Walter Scheel, a Free Democrat, was elected federal president in 1974. Because laissez-faire elements in the Free Democratic Party resisted increases in the government's role in the economy, the Social Democrats were able to achieve little of their program for expanding the welfare state.

The social-liberal coalition came to an end in October 1982 when the Free Democrats, who had been suffering losses in local and regional elections, defected and formed a new coalition with the Christian Democrats. The new chancellor was the veteran Christian Democrat Helmut Kohl. The 1983 Bundestag election yielded sizable gains for the Christian Democrats but heavy losses for their coalition partners, the Free Democrats, many of whose former voters preferred collaboration with the Social Democrats. The Social Democrats also suffered heavy losses, most of which went to the new ecological party, the Greens, who entered the Bundestag only a few years after their movement had formed to protest environmental pollution, sexism, and authoritarianism. In 1984 Richard von Weizsäcker of the Christian Democrats took office as West Germany's sixth federal president.

Faced with mounting dissent at home, the East German government under the leadership of Honecker sought to enhance its legitimacy by seeking further recognition from West Germany. It was therefore buoyed when Chancellor Schmidt paid an often-postponed official visit to the German Democratic Republic in 1981. Schmidt ignored Honecker's demand that Bonn treat East Germans as foreigners and cease to bestow West German citizenship automatically on those who fled to the West.

Nevertheless, after Schmidt's visit East Germany began making it easier for its citizens to visit West Germany. By 1986 nearly 250,000 East Germans were visiting West Germany each year. Only one family member at a time was permitted to go, so that virtually all returned home. The East German government also began granting some of its dissatisfied citizens permission to emigrate to the West. In return, West Germany guaranteed several large Western bank loans to East Germany. In 1987 the East German government realized a long-held ambition when, after many postponements, Honecker was received in Bonn with full state honours, seemingly confirming West Germany's acceptance of the permanence of the German Democratic Republic.

But behind the Honecker government's facade of stability, East Germany was losing its legitimacy in the eyes of the majority of its citizenry. Particularly among younger East Germans, the new opportunities for travel to West Germany produced discontent rather than satisfaction. There, they experienced a much more advanced, consumer-oriented society that provided its citizens with an abundance of far higher-quality goods than were available at home. While in West Germany, they chafed at having to depend materially on Western relatives because their own currency was virtually worthless outside the German Democratic Republic. They also experienced freedom of expression and an open marketplace of ideas and opinions that contrasted sharply with the rigid censorship and repression of deviant views at home. Once these East Germans had traveled or even heard of others' travel, the Berlin Wall and the other border fortifications designed to restrict their movements seemed more onerous than ever. In protest against the East German government's indifference to the damage its outdated industries were inflicting on the environment, a clandestine ecological movement arose, and an underground peace movement opposed the regime's manipulation of the cause of peace for propaganda purposes. Both of these movements found sanctuary in the churches of predominantly Protestant East Germany.

The reunification of Germany

The swift and unexpected downfall of the German Democratic Republic was triggered by the decay of the other communist regimes in eastern Europe and the Soviet Union. The liberalizing reforms of President Mikhail S. Gorbachev in the Soviet Union appalled the Honecker regime, which in desperation was by 1988 forbidding the circulation within East Germany of Soviet publications it viewed as dangerously subversive. The Berlin Wall was in effect breached in the summer of 1989 when a reformist Hungarian government began allowing East Germans to escape to the West through Hungary's newly opened border with Austria. By the fall, thousands of East Germans had followed this route, while thousands of others sought asylum in the West German embassies in Prague and Warsaw to obtain passage to the Federal Republic. Mass demonstrations in the streets of Leipzig and other East German cities defied the authorities and demanded reforms.

In an effort to halt the deterioration of its position, the SED Politburo replaced Honecker in mid-October with another hard-line communist, Egon Krenz. Under Krenz the Politburo sought to end the embarrassing flow of refugees to the West German embassies in neighbouring countries. It intended to allow direct travel to West Germany with a permit issued by the East German government. However, when an official announced on television that the government was permitting travel to West Germany "immediately," crowds gathered at the Berlin Wall demanding to pass into West Berlin. Unprepared, the border guards let them go. In a night of revelry, tens of thousands of East Germans poured through the crossing points in the wall and celebrated their new freedom with rejoicing West Berliners.

Fall of the Berlin Wall

The opening of the Berlin Wall proved fatal for the German Democratic Republic. Ever-larger demonstrations demanded a voice in government for the people, and in mid-November Krenz was replaced by a reform-minded communist, Hans Modrow, who promised free, multiparty elections. When the balloting took place in March 1990 the SED, now renamed the Party of Democratic Socialism (PDS), suffered a crushing defeat. The eastern counterpart of Kohl's Christian Democratic Union, which had pledged a speedy reunification of Germany, emerged as the largest political party in East Germany's first democratically elected People's Chamber. A new East German government headed by Lothar de Maizière, a long-time member of the eastern Christian Democratic Union, and backed initially by a broad coalition including the Social Democrats and Free Democrats, began negotiations for a treaty of unification. A surging tide of refugees from East to West Germany that threatened to cripple the German Democratic Republic added urgency to those negotiations. In July that tide was somewhat stemmed by a monetary union of the two Germanies that gave East Germans the hard currency of the Federal Republic.

The final obstacle to reunification was removed in July 1990 when Kohl prevailed upon Gorbachev to drop his objections to a unified Germany within the NATO alliance in return for sizable (West) German financial aid to the Soviet Union. A unification treaty was ratified by the Bundestag and the People's Chamber in September and went into effect on Oct. 3, 1990. The German Democratic Republic joined the Federal Republic as five additional *Länder,* and the two parts of divided Berlin became one *Land.* (The five new *Länder* were Brandenburg, Mecklenburg–West Pomerania, Saxony, Saxony-Anhalt, and Thuringia.) After 45 years of division, Germany was once again a united nation.

In December 1990 the first all-German free election since the 1930s conferred an expanded majority on the coalition government led by Kohl. (H.A.T.)

Difficulties attending unification

Unification soon faced a series of difficulties. Like most of Europe, Germany in the 1990s confronted increased global competition and stubborn unemployment, especially in its traditional industrial sector. However, it also faced the staggering expense of unifying the former east and west. The monetary union of East and West Germany, which had involved a one-to-one exchange of East German for West German marks, priced most East German goods and services well above their market value. Under these conditions, only a handful of eastern firms could compete on the world market. As a result the former East German economy collapsed, leaving hundreds of thousands of easterners facing unemployment. Moreover, the east became heavily dependent on subsidies from the federal government. Meanwhile, the infrastructure—roads, rail lines, and telephones—required massive capital investment. Unemployment, social dislocation, and disappointment continued to haunt the former east into the 21st century.

Many easterners resented economic hardship and perceived western domination, while many westerners chafed at the heavy taxation that financed the rebuilding of the east. The PDS became the political voice of eastern discontent. Meanwhile, there was the problem of resolving the legacies of 40 years of dictatorship. East Germany's large and effective State Security Police (Staatssicherheitspolizei, or Stasi) had employed a wide network of professional and amateur informants. When the files of this organization were made public, eastern Germans discovered that many of their most prominent citizens, as well as some of their friends, neighbours, and even family members, had been on the Stasi payroll.

Despite the problems attending unification and a series of scandals in his own party, Kohl won a narrow victory in the 1994 elections. Earlier that year, Roman Herzog had assumed the presidency. In 1996 Kohl became the longest-serving German chancellor since Bismarck. Nevertheless, his popularity was clearly ebbing. Some members of Kohl's party hoped that he would step aside in favour of a new candidate in the 1998 elections. Instead, the chancellor ran again and lost to Gerhard Schröder, the pragmatic and photogenic Social Democratic candidate, who formed a coalition with the Greens.

Schröder's government was largely concerned with reducing Germany's high unemployment and reviving the country's stagnant economy. In 1999 Schröder deployed troops to Kosovo to stop ethnic cleansing there, and he also sent forces to Afghanistan in 2001 as part of an international effort against terrorism after the attacks against the United States on September 11 of that year. However, despite the implementation of various domestic economic reforms, by 2002 unemployment remained stubbornly high, and the worldwide economic downturn continued to dampen the return to economic growth. Nonetheless, aided by his response to historic floods in Germany and his opposition to U.S. military action against Iraq, Schröder's SPD-Green coalition narrowly won reelection in September 2002.

At the start of the new millennium, Germany remained the most powerful country in Europe. For more than 50 years, Germans had played an important role in the creation of European institutions. German support remained essential to the success both of the European Union's ambitious program of economic and political integration and of European efforts to supplement or replace Cold War institutions such as NATO with new security arrangements. (J.J.Sh./Ed.)

For later developments in the history of Germany, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 912, 921, 923, 961, 963, and 972, and the *Index.*

BIBLIOGRAPHY

Geography. A simple but useful portrayal of Germany in maps is BERNHARD SCHÄFERS *et al., The State of Germany Atlas* (1998; originally published in German, 1997). W. TIETZE *et al.* (eds.), *Geographie Deutschlands: Bundesrepublik Deutschland* (1990), is a massive, comprehensive, and authoritative geographic survey. Other useful sources include ELMAR KULKE, *Wirtschaftsgeographie Deutschlands* (1998), on economic geography; and HERBERT LIEDTKE and JOACHIM MARCINEK (eds.), *Physische Geographie Deutschlands* (1994), on physical geography. LUIGI BARZINI, *The Impossible Europeans* (1983), discusses the nation in the context of the European federation. ERIC OWEN SMITH, *The German Economy* (1994), provides an overview of West and East Germany's economies as well as the economy of united Germany. Sources of recent information on Germany include the periodicals *Country Profile: Germany* (annual), *Country Report: Germany* (quarterly), and *OECD Economic Surveys: Germany* (annual). The current economic policies and perspectives of the German federal government are outlined in annual economic reports, such as *The German Government's Economic Report for 2000: Economic and Financial Reports: Creating Jobs—Strengthening Germany's Economic Future* (2000). Recent publications that provide an overview of population issues in Germany and Europe include RAINER MÜNZ and MYRON WEINER (eds.), *Migrants, Refugees, and Foreign Policy: U.S. and German Policies Toward Countries of Origin* (1997); and HEINZ FASSMANN and RAINER MÜNZ (eds.), *European Migration in the Late Twentieth Century: Historical Patterns, Actual Trends, and Social Implications* (1994). (Wm.H.B.)

The administrative and political structure of former West Germany are covered in PETER J. KATZENSTEIN, *Policy and Politics in West Germany: The Growth of a Semi-Sovereign State* (1987); and IAN DERBYSHIRE, *Politics in West Germany,* rev. 2nd ed. (1991). Sources on intellectual life, arts, and culture in former West Germany include GORDON A. CRAIG, *The Germans* (1982, reprinted 1991); K. STUART PARKES, *Writers and Politics in West Germany* (1986); and JOHN SANDFORD, *The New German Cinema* (1980, reissued 1982). (G.H.K.)

Among the comprehensive works on former East Germany are STEPHEN B. BURANT (ed.), *East Germany: A Country Study,* 3rd ed. (1988); KARL ECKART, *DDR,* 3rd. ed. (1989); H. KOHL, J. MARCINEK, and B. NITZ, *Geography of the German Democratic Republic* (1986; originally published in German, 1978); HENRY KRISCH, *The German Democratic Republic: The Search for Identity* (1985); and AKADEMIE DER WISSENSCHAFTEN DER DDR, *Information GDR: The Comprehensive and Authoritative Reference Source of the German Democratic Republic,* 2 vol. (1989). DAVID CHILDS, THOMAS A. BAYLIS, and MARILYN RUESCHEMEYER (eds.), *East Germany in Comparative Perspective* (1989), covers a broad spectrum of topics. MIKE DENNIS, *German Democratic Republic: Politics, Economics, and Society* (1988), surveys the administrative, political, and social structure of former East Germany. (T.H.El.)

Several works compare the two former states or treat them together. TERENCE PRITTIE, *My Germans, 1933–1983* (1983), explores social life and customs and national characteristics; as does HARRY G. SHAFFER, *Women in the Two Germanies: A Comparative Study of a Socialist and a Non-Socialist Society* (1981). LESLIE LIPSCHUTZ and DONOGH MCDONALD (eds.), *German Unification: Economic Issues* (1990), deals with the economic ramifications of political unification. ERNEST D. PLOCK, *The Basic Treaty and the Evolution of East-West German Relations* (1986), discusses relations between the two countries. Postwar literature is the subject of LOWELL A. BANGERTER, *German Writing Since 1945: A Critical Survey* (1988). CELIA APPLEGATE, *A Nation of Provincials: The German Idea of Heimat* (1990), considers the pervading German attachment to local place. (G.H.K./T.H.El.)

History. *Ancient Germany:* HERBERT SCHUTZ, *The Prehistory of Germanic Europe* (1983), offers a good, well-illustrated general account of the early period. TIM CORNELL and JOHN MATTHEWS, *Atlas of the Roman World* (1982, reissued 1987), employs maps and illustrations to clarify the social and economic background of Roman-German interaction. MALCOLM TODD, *The Northern Barbarians, 100 BC–AD 300,* rev. ed. (1987), is another useful source on prehistory and early history. EDWARD JAMES, *The Franks* (1988, reissued 1991), discusses one of the more important Germanic peoples. (P.J.He.)

Merovingians and Carolingians: PATRICK J. GEARY, *Before France and Germany: The Creation and Transformation of the Merovingian World* (1988), examines the period from the perspective of the unity of late antiquity. ROSAMOND MCKITTERICK, *The Frankish Kingdoms Under the Carolingians, 751–987* (1983), presents a general survey of Carolingian Europe. TIMOTHY REUTER (ed. and trans.), *The Medieval Nobility: Studies on the Ruling Classes of France and Germany from the Sixth to the Twelfth Century* (1979), contains important essays on early German and Frankish society. (P.J.G.)

Medieval Germany to 1250: G. BARRACLOUGH, *The Origins of Modern Germany*, 3rd ed. (1988), is a good analysis of medieval German history. General surveys include HORST FUHRMANN, *Germany in the High Middle Ages, c. 1050–1200* (1986; originally published in German, 1978), covering such fundamentals as economy and cultural horizons; ALFRED HAVERKAMP, *Medieval Germany, 1056–1273*, 2nd ed. (1992; originally published in German, 1984), providing an informative though mainly German bibliography; and TIMOTHY REUTER, *Germany in the Early Middle Ages, c. 800–1056* (1991), presenting a good synthesis of more recent scholarship. K.J. LEYSER, *Rule and Conflict in an Early Medieval Society: Ottonian Saxony* (1979, reissued 1989), explores the main forces of development of the Saxon empire and its society from 900 to 1024, and *Medieval Germany and Its Neighbours, 900–1250* (1982), studies warfare, the nobility, and German-Byzantine and German-English relations. BENJAMIN ARNOLD, *German Knighthood, 1050–1300* (1985), is a fine study of knights and ministeriales. (K.J.L.)

From 1250 to 1493: A valuable source on this period is F.R.H. DU BOULAY, *Germany in the Later Middle Ages* (1983). HERMANN KELLENBENZ, *Von den Anfängen bis zum Ende des 18. Jahrhunderts* (1977), vol. 1 of *Deutsche Wirtschaftsgeschichte*, surveys medieval and early modern economic history. The collection of GERHART HOFFMEISTER (ed.), *The Renaissance and Reformation in Germany: An Introduction* (1977), offers a fine overview of intellectual life; and on the extraordinary flowering of German sculpture there are two masterly works by MICHAEL BAXANDALL, *South German Sculpture, 1480–1530* (1974), and *The Limewood Sculptors of Renaissance Germany* (1980). (L.G.D.)

From 1493 to c. 1760: The best guide to this period is THOMAS A. BRADY, JR., HEIKO A. OBERMAN, and JAMES D. TRACY (eds.), *Handbook of European History, 1400–1600: Late Middle Ages, Renaissance, and Reformation*, 2 vols. (1994–95). BOB SCRIBNER and SHEILAGH OGILVIE (eds.), *Germany: A New Social and Economic History*, 2 vols. (1996), is a good collection on the society and economy of the Renaissance and Reformation. HAJO HOLBORN, *A History of Modern Germany: The Reformation* (1959, reprinted 1982), is a comprehensive general account, with emphasis on political events. PETER BLICKLE, *The Revolution of 1525: The German Peasants' War from a New Perspective* (1981, reissued 1985; originally published in German, 1975), stresses the common grievances that brought peasants and the urban poor together in a widespread uprising. THOMAS A. BRADY, JR., *The Politics of the Reformation in Germany: Jacob Sturm (1489–1553) of Strasbourg* (1997), tells the story of Reformation politics from the perspective of an important Protestant statesman. MERRY E. WIESNER, *Working Women in Renaissance Germany* (1986), offers a pioneering examination of the changing situation of women engaged in trades and professions between 1500 and 1600. GEOFFREY PARKER (ed.), *The Thirty Years' War*, 2nd rev. ed. (1997), surveys the military, political, social, and cultural aspects of the war. JOHN G. GAGLIARDO, *Reich and Nation: The Holy Roman Empire as Idea and Reality, 1763–1806* (1980), is an examination of the empire's efforts and failure to survive in the age of the French Revolution. R.J.W. EVANS, *The Making of the Habsburg Monarchy, 1550–1700: An Interpretation* (1979, reprinted 1991), analyzes the events leading to the construction of the Austrian empire. RUDOLF VIERHAUS, *Germany in the Age of Absolutism* (1988; originally published in German, 1978), discusses the German principalities in the late 17th and 18th centuries. (Ge.St./ J.J.Sh.)

From c. 1760 to 1871: A good survey of the period as a whole is JAMES J. SHEEHAN, *German History, 1770–1866* (1989, reissued 1993). DAVID BLACKBOURN, *The Long Nineteenth Century: A History of Germany, 1780–1918* (also published as *Fontana History of Germany, 1780–1918: The Long Nineteenth Century*, 1997), covers Germany's emergence as a nation-state. The best account of Germany's place in the international system is PAUL W. SCHROEDER, *The Transformation of European Politics, 1763–1848* (1994, reissued 1996). The most comprehensive source on German social history is the monumental HANS-ULRICH WEHLER, *Deutsche Gesellschaftsgeschichte*, (1987–). C.B.A. BEHRENS, *Society, Government, and the Enlightenment: The Experiences of Eighteenth-Century France and Prussia* (1985), is a comparative history of the events and developments that led to revolution in France and stabilization in Germany. T.C.W. BLANNING, *The French Revolution in Germany: Occupation and Resistance in the Rhineland, 1792–1802* (1983), is a re-examination of this period of German history. ENNO KRAEHE, *Metternich's German Policy*, 2 vol. (1963–83), describes the War of Liberation; while HAROLD NICOLSON, *The Congress of Vienna: A Study in Allied Unity, 1812–1822* (1946, reissued 2000), deals with the reconstruction of Europe. A balanced and up-to-date overview of the period of the German Confederation is THOMAS NIPPERDEY, *Germany from Napoleon to Bismarck, 1800–1866* (1996; originally published in German, 1983). JONATHAN SPERBER, *Rhineland Radicals: The Democratic Movement and the Revolution of 1848–1849* (1991, reprinted 1993), provides a good analysis of the revolution's social origins and political meaning in western Germany. OTTO PFLANZE, *Bismarck and the Development of Germany*, 2nd ed., 3 vol. (1990), portrays the architect of the new Germany and his impact on the European balance of power. (T.S.H./J.J.Sh./Ed.)

From 1871 to 1918: General historical surveys include GORDON A. CRAIG, *Germany, 1866–1945* (1978, reissued 1981), which is particularly good on politics and culture; and HANS-ULRICH WEHLER, *The German Empire, 1871–1918* (1985, reissued 1997; originally published in German, 1973), which stresses the continuity between the empire and the Nazi era. DAVID BLACKBOURN and GEOFF ELEY, *The Peculiarities of German History: Bourgeois Society and Politics in Nineteenth-Century Germany* (1984), explores the connection between economics and politics in the German Empire. MARGARET LAVINIA ANDERSON, *Practicing Democracy: Elections and Political Culture in Imperial Germany* (2000), examines the relationship between political participation and authority from 1867 to 1914. The history of German socialism is detailed in GARY P. STEENSON, *"Not One Man! Not One Penny!": German Social Democracy, 1863–1914* (1981), a comprehensive treatment of the Social Democratic Party from its origins to the war. ROGER CHICKERING (ed.), *Imperial Germany: A Historiographical Companion* (1996), is a useful guide to further reading. A valuable source on developments leading to World War I is V.R. BERGHAHN, *Germany and the Approach of War in 1914*, 2nd ed. (1993). A good introduction to Germany at war is ROGER CHICKERING, *Imperial Germany and the Great War, 1914–1918* (1998). JÜRGEN KOCKA, *Facing Total War: German Society, 1914–1918* (1984; originally published in German, 1973), examines the effects of World War I on different social groups. UTE FREVERT, *Women in German History: From Bourgeois Emancipation to Sexual Liberation* (1989, reissued 1993; originally published in German, 1986), is a fine survey of this important area of research. (Ke.B./ J.J.Sh.)

From 1918 to 1945: DIETRICH ORLOW, *A History of Modern Germany: 1871 to Present*, 4th ed. (1999), offers a good survey of Germany's troubled 20th century. A.J. RYDER, *The German Revolution of 1918: A Study of German Socialism in War and Revolt* (1967), treats both the democratic and revolutionary socialists in 1918 and 1919 along with their conflicting visions for change. HANS MOMMSEN, *The Rise and Fall of Weimar Democracy*, trans. by ELBORG FORSTER and LARRY EUGENE JONES (1996; originally published in German, 1989), is a fine synthesis of the republic's tragic history. The definitive study of the German inflation is GERALD D. FELDMAN, *The Great Disorder: Politics, Economics, and Society in the German Inflation, 1914–1924* (1993, reissued 1997). PETER GAY, *Weimar Culture: The Outsider as Insider* (1968, reprinted 1988), provides a stimulating introduction to the cultural and intellectual brilliance of Germany during the 1920s. An excellent guide to further reading on Weimar culture is ANTON KAES, MARTIN JAY, and EDWARD DIMENDBERG (eds.), *The Weimar Republic Sourcebook* (1994). IAN KERSHAW, *Hitler, 1889–1936: Hubris* (1998, reissued 2000), and *Hitler, 1936–45: Nemesis* (2000), together constitute the definitive biography of the dictator. A comprehensive study of National Socialism emphasizing the totalitarian nature of the regime is available in KARL DIETRICH BRACHER, *The German Dictatorship: The Origins, Structure, and Effects of National Socialism* (1970, reissued 1989; originally published in German, 1969). One of the best single-volume treatments of World War II is GERHARD L. WEINBERG, *A World at Arms: A Global History of World War II* (1994). RAUL HILBERG, *The Destruction of the European Jews*, rev. ed., 3 vol. (1985), remains the standard and most comprehensive study of the efforts by the Nazis to exterminate the Jewish people; SAUL FRIEDLÄNDER, *Nazi Germany and the Jews* (1997–), is an important newer study. EARL R. BECK, *Under the Bombs: The German Home Front, 1942–1945* (1986), explores German civilian life during the war. (K.A.Sc./J.J.Sh./Ed.)

After 1945: Surveys of the history of both German states include HENRY ASHBY TURNER, JR., *The Two Germanies Since 1945* (1987); and A. JAMES MCADAMS, *Germany Divided: From the Wall to Reunification* (1993). V.R. BERGHAHN, *Modern Germany: Society, Economy, and Politics in the Twentieth Century*, 2nd ed. (1987, reissued 1993), is a comparative history. A more

detailed account of West Germany's development is DENNIS L. BARK and DAVID R. GRESS, *A History of West Germany*, 2nd ed., 2 vol. (1993). PETER H. MERKL, *The Origin of the West German Republic* (1963, reprinted 1982), focuses on West Germany's beginnings; and ROGER MORGAN, *The United States and West Germany, 1945–1973: A Study in Alliance Politics* (1974), treats its relations with the United States. JEREMY LEAMAN, *The Political Economy of West Germany, 1945–85: An Introduction* (1988), surveys the West German economy. A good account of the Soviet role in East Germany is NORMAN M. NAIMARK, *The Russians in Germany: A History of the Soviet Zone of Occupation, 1945–1949* (1995). Other sources on the history of East Germany include MARY FULBROOK, *Anatomy of a Dictatorship: Inside the GDR, 1949–1989* (1995, reissued 1997); and DAVID CHILDS (ed.), *Honecker's Germany* (1985). The best single-volume treatment of the collapse of East Germany is CHARLES S. MAIER, *Dissolution: The Crisis of Communism and the End of East Germany* (1997, reissued 1999). KONRAD H. JARAUSCH, *The Rush to German Unity* (1994), is also useful.

(H.A.T./J.J.Sh.)

Globalization and Culture

Globalization is the process by which the experience of everyday life, marked by the diffusion of commodities and ideas, is becoming standardized around the world. An extreme interpretation of this process, often referred to as globalism, sees advanced capitalism, boosted by wireless and Internet communications and electronic business transactions, destroying local traditions and regional distinctions, creating in their place a homogenized world culture. According to this view, human experience everywhere is in jeopardy of becoming essentially the same. This appears, however, to be an overstatement of the phenomenon. Although homogenizing influences do indeed exist, people are far from creating a single overarching world culture. The actual process of globalization has been fitful, chaotic, and slow.

This article is divided into the following sections:

Emergence of global subcultures 133
"Davos" culture
The international "faculty club"
Nongovernmental organizations
Transnational workers
The persistence of local culture 133
Experiencing globalization 134
The collapse of time and space
The standardization of experience
Political consequences of globalization 136
Challenges to national sovereignty and identity
Anti-globalism movements and the Internet
The illusion of global culture 137
Localized responses
Borrowing and "translating" popular culture
Subjectivity of meaning—the case of *Titanic*
The ties that still bind 137
Bibliography 137

EMERGENCE OF GLOBAL SUBCULTURES

Some observers argue that a rudimentary version of world culture is already taking shape among certain individuals who share similar values, aspirations, or lifestyles. The result is a collection of elite groups whose unifying ideals transcend geographical limitations.

"Davos" culture. One such cadre, according to political scientist Samuel Huntington in *The Clash of Civilizations* (1998), comprises an elite group of highly educated people who operate in the rarefied domains of international finance, media, and diplomacy. Named after the Swiss town that hosts annual meetings of the World Economic Forum, these "Davos" insiders share common beliefs about individualism, democracy, and market economics. They are said to follow a recognizable lifestyle, are instantly identifiable anywhere in the world, and feel more comfortable in each other's presence than they do among their less-sophisticated compatriots.

The international "faculty club." The globalization of cultural subgroups is not limited to the upper classes, however. Expanding on the concept of Davos culture, sociologist Peter L. Berger observes that the globalization of Euro-American academic agendas and lifestyles has created a worldwide "faculty club"—an international network of people who share similar values, attitudes, and research goals. Although they are not as wealthy or privileged as their Davos counterparts, members of this international faculty club wield tremendous influence through their association with educational institutions worldwide and have been instrumental in promoting feminism, environmentalism, and human rights as global issues. Berger cites the antismoking movement as a case in point: the movement began as a singular North American preoccupation in the 1970s and subsequently spread to other parts of the world, traveling along the contours of academe's global network.

Nongovernmental organizations. Another global subgroup comprises "cosmopolitans" who nurture an intellectual appreciation for local cultures. As pointed out by Swedish anthropologist Ulf Hannerz, this group advocates a view of global culture based not on the "replication of uniformity" but on the "organization of diversity." Often promoting this view are nongovernmental organizations (NGOs) that lead efforts to preserve cultural diversity in the developing world. By the turn of the 21st century, institutions such as Cultural Survival were operating on a world scale, drawing attention to indigenous groups who are encouraged to see themselves as "first peoples"—a new global designation emphasizing common experiences of exploitation among indigenous inhabitants of all lands. By sharpening such identities, these NGOs have globalized the movement to preserve indigenous world cultures.

Transnational workers. Another group stems from the rise of a transnational workforce. Indian-born Arjun Appadurai, an American anthropologist, studied English-speaking professionals who trace their origins to South Asia but who live and work elsewhere. They circulate in a social world that has multiple home bases, and they have gained access to a unique network of individuals and opportunities. For example, many software engineers and Internet entrepreneurs who live and work in Silicon Valley, California, in the United States, maintain homes in—and strong social ties to—Indian states such as Maharashtra and Punjab.

THE PERSISTENCE OF LOCAL CULTURE

Underlying these various visions of globalization is a reluctance to define exactly what is meant by the term *culture.* During most of the 20th century, anthropologists defined culture as a shared set of beliefs, customs, and ideas that held people together in recognizable, self-identified groups. Scholars in many disciplines challenged this notion of cultural coherence, especially as it became evident that members of close-knit groups held radically different visions of their social worlds. As a result, many social scientists now treat culture as a set of ideas, attributes, and expectations that change as people react to changing circumstances. Indeed, at the turn of the 21st century, the collapse of barriers enforced by Soviet socialism and the rise of electronic commerce have increased the perceived speed of social change everywhere.

Place and location still matter

The term *local culture* is commonly used to characterize the experience of everyday life in specific, identifiable localities. It reflects ordinary peoples' feelings of appropriateness, comfort, and correctness—attributes that define personal preferences and changing tastes. Given the strength of local cultures, it is difficult to argue that an overarching global culture actually exists. Jet-setting sophisticates may feel comfortable operating in a global network disengaged from specific localities, but these people constitute a very small minority; their numbers are insufficient to sustain a coherent cultural system. It is more important to ask where these global operators maintain their families, what kind of kinship networks they rely upon, if any, and if theirs is a transitory lifestyle or a permanent condition. For most people, place and locality still matter. Even the transnational workers discussed by Appadurai are rooted in local communities bound by common perceptions of what represents an appropriate and fulfilling lifestyle.

EXPERIENCING GLOBALIZATION

Research on globalization has shown that it is not an omnipotent, unidirectional force leveling everything in its path. Because a global culture does not exist, any search for it will be futile. It is more fruitful to focus on particular aspects of life that are indeed affected by the globalizing process.

The collapse of time and space. The breakdown of time and space is best illustrated by the influential "global village" thesis by Marshall McLuhan in *Gutenberg Galaxy* (1962). Instantaneous communication, predicted McLuhan, would soon destroy geographically based power imbalances and create a global village. Later, geographer David Harvey argued that the postmodern condition is characterized by a "time-space compression" that arises from inexpensive air travel and the ever-present use of telephones, fax, and, more recently, E-mail.

There can be little doubt that people perceive the world today as a smaller place than it appeared to their grandparents. In the 1960s and '70s immigrant workers in London relied on postal systems and personally delivered letters to send news back to their home villages in India, China, and elsewhere; it could take two months to receive a reply. The telephone was not an option, even in dire emergencies. By the late 1990s, the grandchildren of these first-generation migrants were carrying cellular phones that linked them instantaneously to cousins in Calcutta, Singapore, or Shanghai.

Live global news coverage

McLuhan's notion of the global village presupposed the worldwide spread of television, which brings distant events into the homes of viewers everywhere. Building on this concept, McLuhan claimed that accelerated communications produce an "implosion" of personal experience—that is, distant events are brought to the immediate attention of people halfway around the world.

The spectacular growth of Cable News Network (CNN) is a case in point. CNN became an icon of globalization through its U.S.-style news programming broadcast throughout the world, 24 hours a day. Live coverage of the fall of the Berlin Wall in 1989 and the Gulf War in 1991 illustrated the power of television to influence world opinion. Some governments have responded to such advances by attempting to restrict international broadcasting, but satellite technology makes these restrictions increasingly unenforceable.

The standardization of experience. *Travel.* Since the mid-1960s, the cost of international flights has declined, and foreign travel has become a routine experience for millions of middle- and working-class people. Diplomats, businesspeople, and ordinary tourists can feel "at home" in any city, anywhere in the world. Foreign travel need no longer involve the challenge of adapting to unfamiliar food and living arrangements. Western-style beds, toilets, showers, fitness centres, and restaurants now constitute the global standard, and CNN has been an essential feature of the standardized hotel experience since at least the 1990s. These developments are linked to the technology of climate control. In fact, the very idea of routine global travel was inconceivable prior to the universalization of air-conditioning. An experience of this nature would have been nearly impossible in the 1960s, when the weather, aroma, and noise of the local society pervaded one's hotel room.

Clothing. Modes of dress can disguise an amazing array of cultural diversity behind a facade of uniformity. The man's business suit, with coloured tie and buttoned shirt, is now "universal" in the sense that it is worn just about everywhere—even in parts of the world that work to avoid global trends, such as Saudi Arabia and Iran. Iranian parliamentarians wear the "Western" suit but forgo the tie, while Saudi diplomats alternate "traditional" Bedouin robes with tailored business suits, depending upon the occasion. At the turn of the 21st century, North Korea and Afghanistan appeared to be among the few societies holding out against these globalizing trends.

The emergence of women's "power suits" in the 1980s signified another form of global conformity. Stylized trouser-suits, with silk scarves and colourful blouses (analogues of the male business suit), are now worldwide symbols of modernity, independence, and competence. Moreover, the export of used clothing from Western countries to developing nations has accelerated the adoption of Western-style dress around the world.

Entertainment. The power of media conglomerates and the ubiquity of entertainment programming has globalized television's impact and made it a logical target for accusations of cultural imperialism. Critics cite a 1999 anthropological study that linked the appearance of anorexia in Fiji to the popularity of American television programs, notably *Melrose Place* and *Beverly Hills 90210.* Both series feature the exploits of slender young actresses who, it was claimed, led Fijian women (who are typically fuller-figured) to question indigenous notions of the ideal body.

Anti-globalism activists contend that American television shows have corrosive effects on local cultures by highlighting Western notions of beauty, individualism, and sexuality. Although many of the titles exported are considered second-tier shows in America, there is no dispute that these programs are part of the daily fare for viewers around the world. Television access is widespread, even if receivers are not present in every household. In the small towns of Guatemala, the villages of Jiangxi province in China, or the hill settlements of Borneo, for instance, one television set—often a satellite system powered by a gasoline generator—may serve two or three dozen viewers, each paying a small fee. Collective viewing in bars, restaurants, and teahouses was common during the early stages of television broadcasting in Indonesia, Japan, Kenya, and many other countries. By the 1980s, video-viewing parlours had become ubiquitous in all but the poorest regions of the globe.

Live sports programs continue to draw the largest global audiences. The 1998 World Cup football (soccer) final between Brazil and France was watched by an estimated two billion people. After the 1992 Olympic Games, when the American "Dream Team" of National Basketball Association (NBA) stars electrified viewers who had never seen the sport played to U.S. professional standards, NBA games were broadcast in Australia, Israel, Japan, China, Germany, and Britain.

Hollywood has had a similar influence, much to the chagrin of some countries. In early 2000, Canadian government regulators ordered the Canadian Broadcasting Corporation (CBC) to reduce the showing of Hollywood films during prime time and feature more Canadian-made programming. CBC executives protested that their viewers would stop watching Canadian television stations and turn to satellite reception for international entertainment. Such objections were well grounded, given that 79 percent of English-speaking Canadians named a U.S. program when asked to identify their favourite television show in 1998.

Popularity of the Indian film industry

Hollywood, however, does not hold a monopoly on entertainment programming. The world's most prolific film industry is in Bombay, India ("Bollywood"), where as many as 1,000 feature films are produced annually in all of India's major languages. Primarily love stories with heavy doses of singing and dancing, Bollywood movies are popular throughout Southeast Asia and the Middle East. State censors in Islāmic countries often find the modest dress and subdued sexuality of Indian film stars acceptable for their audiences. Although the local appeal of Bollywood movies remains strong, exposure to Hollywood films such as *Jurassic Park* (1993) and *Speed* (1994) has caused young Indian moviegoers to develop an appreciation for the special effects and computer graphics that are hallmarks of most American films.

Food. It is not mass media but food that is the oldest global carrier of culture. In fact, food has always been a driving force for globalization, especially during earlier phases of European trade and colonial expansion. The hot red pepper was introduced to the Spanish court by Christopher Columbus in 1493. It spread rapidly throughout the colonial world, transforming cuisines and farming practices in Africa, Asia, and the Middle East. It might be difficult to imagine Korean cuisine without red pepper paste or Szechuan food without its fiery hot sauce, but both are relatively recent innovations—probably from the 17th century. Other New World crops, such as corn (maize), cassava, sweet potatoes, and peanuts (groundnuts), were responsible for agricultural revolutions in Asia and Africa,

opening up terrain that had previously been unproductive.

One century after the sweet potato was introduced into south China, it had become a dominant crop and was largely responsible for a population explosion that created what today is called Cantonese culture. It is the sweet potato, not the more celebrated white rice, that sustained generations of southern Chinese farmers. These are the experiences that cause cultural meaning to be attached to particular foods. Today the descendants of Cantonese, Hokkien, and Hakka pioneers disdain the sweet potato as a "poverty food" that conjures up images of past hardships. In Taiwan, by contrast, independence activists (affluent members of the new Taiwanese middle class) have embraced the sweet potato as an emblem of identity, reviving old recipes and celebrating their cultural distinctions from "rice-eating mainlanders."

The global food trade

While the global distribution of foods originated with the pursuit of exotic spices (such as black pepper, cinnamon, and cloves), contemporary food trading features more prosaic commodities such as soybeans, bananas, and oranges. Green beans are now grown in Burkina Faso in Central Africa and shipped by express air cargo to Paris, where they end up on the plates of diners in five-star restaurants. This particular exchange system is based on a "nontraditional" crop that was not grown in Burkina Faso until the mid-1990s, when the World Bank encouraged its cultivation as a means of promoting economic development. The country soon became Africa's second-largest exporter of green beans. Central African farmers consequently find themselves in direct competition with other "counter-season" producers of green beans in Brazil and Florida. African bananas, Chilean grapes, and California oranges have helped to transform consumer expectations about the availability of fresh produce everywhere in the world.

The average daily diet has also undergone tremendous change, with all nations converging on a diet high in meat, dairy products, and processed sugars. Correlating closely to a worldwide rise in affluence, the new "global diet" is not necessarily a beneficial trend, as it can increase the risk of obesity and diabetes. Now viewed as a global health threat, obesity has been dubbed "globesity" by the World Health Organization. To many observers, the homogenization of human diet appears to be unstoppable. Vegetarians, environmental activists, and organic food enthusiasts have organized rearguard actions to reintroduce "traditional" and more wholesome dietary practices, but these efforts are concentrated among educated elites in industrial nations.

Western corporations are often blamed for these dietary trends. McDonald's, KFC (Kentucky Fried Chicken), and Coca-Cola are primary targets of anti-globalism demonstrators (who are themselves organized into global networks, via the Internet). Food and beverage companies attract attention because they cater to the most elemental form of human consumption. We are what we eat; when diet changes, notions of national and ethnic identity are affected. McDonald's has become a symbol of globalism for obvious reasons: On an average day in 2001, the company served nearly 45 million customers at more than 25,000 restaurants in 120 countries. It succeeds in part by adjusting its menu to local needs. In India, for example, no beef products are sold.

The influences of fast food restaurants are not limited to dietary changes. In Japan, for instance, using one's hands to eat prepared foods was considered a gross breach of etiquette. The popularization of McDonald's hamburgers has had such a dramatic impact on popular etiquette that it is now common to see Tokyo commuters eating in public, without chopsticks or spoons. In late-Soviet Russia, rudeness had become a high art form among service personnel. Today customers expect polite, friendly service when they visit Moscow restaurants—a social revolution initiated by McDonald's and its employee indoctrination programs.

The social atmosphere in colonial Hong Kong of the 1960s was anything but genteel. Cashing a check, boarding a bus, or buying a train ticket required brute force. When McDonald's opened in 1975, customers clumped around the cash registers, shouting orders and waving money over the heads of people in front of them. McDonald's responded by introducing queue monitors—young women who channeled customers into orderly lines. Queuing subsequently became a hallmark of Hong Kong's cosmopolitan, middle-class culture. Older residents credit McDonald's for introducing the queue, a critical element in this social transition. In some cultures of Asia, Latin America, and Europe, another innovation was the provision of clean toilets and washrooms. McDonald's was instrumental in setting new cleanliness standards (thereby raising consumer expectations) in cities that had never offered public facilities.



Growing up with fast food: young girls attending a birthday party at a McDonald's restaurant, Shenzhen, China.
Michael S. Yamashita—Corbis

The introduction of fast food has been particularly influential on children who are often the direct targets of television advertising. The symbolic impact of fast food makes it a powerful force for dietary and social change, and a meal at these restaurants will introduce practices that younger consumers may not experience at home—most notably, the chance to choose one's own food. The concept of personal choice is symbolic of Western consumer culture. Visits to McDonald's and KFC have become signal events for children who approach fast food restaurants with a heady sense of empowerment.

Religion. Central to Huntington's thesis in *The Clash of Civilizations* is the assumption that the post-Cold War world would regroup into regional alliances based on religious beliefs and historical attachments to various "civilizations." Identifying three prominent groupings—Western Christianity (Catholicism and Protestantism), Orthodox Christianity (Russian and Greek), and Islām, with additional influence from Hinduism and Confucianism—he predicted that the progress of globalization would be severely constrained by religio-political barriers. The result would be a "multipolar world." This view is quite different from the prophesies of a standardized, homogenized global culture.

Diffusion of religious and cultural systems

There is, however, considerable ethnographic evidence, gathered by anthropologists and sociologists, that refutes this model of civilizational clash and points instead to a rapid diffusion of religious and cultural systems throughout the world. Islām is an excellent case in point, given that it constitutes one of the fastest-growing religions in the United States, France, and Germany—supposed bastions of the Christian West. At the turn of the 21st century, entire districts (*arrondisements*) of Paris were dominated by Muslims, the majority of them French citizens born and reared in France. Thirty-five percent of students in the suburban Dearborn, Mich., public school system were Muslim in 2001; the provision of *halāl* ("lawful" under Islām) meals at lunchtime became a hot issue in local politics. Meanwhile Muslims of Turkish origin constituted the fastest-growing sector of Berlin's population, and, in northern England, the old industrial cities of Bradford and Newcastle were revitalized by descendants of Pakistani and Indian Muslims who immigrated during the 1950s and '60s.

From its inception, Christianity has been an aggressively proselytizing religion with a globalizing agenda. Indeed, the Roman Catholic church was arguably the first global

institution, spreading rapidly throughout the European colonial world and beyond. Today, perhaps the fastest-growing religion is evangelical Christianity. Stressing the individual's personal experience of divinity, evangelism has wide appeal in developing regions such as Latin America and sub-Saharan Africa, presenting serious challenges to established Catholic churches. Following the collapse of Soviet power in 1991, the Russian Orthodox church began the process of rebuilding after more than seven decades of repression. At the same time, evangelical missionaries from the United States and Europe shifted much of their attention from Latin America and Africa to Russia, alarming Russian Orthodox leaders. By 1997, under pressure from Orthodox clergy, the Russian government promoted legislation to restrict the activities of religious organizations that had operated in Russia for less than 15 years, effectively banning Western evangelical missionaries. The debate over Russian religious unity continues, however, and such legislation could have little long-term effect.

Sociologists such as Berger confirm the resurgence, since the late 20th century, of conservative religion among faiths such as Islām, Hinduism, Buddhism, and even Shintō (in Japan) and Sikhism (in India). The social and political connotations of these conservative upsurges are unique to each culture and religion. According to Berger, evangelicalism can be seen as a carrier of modernization: its emphasis on the Bible encourages literacy, while involvement in church activities can teach administrative skills that are applicable to work environments.

POLITICAL CONSEQUENCES OF GLOBALIZATION

Challenges to national sovereignty and identity. Anti-globalism activists often depict the McDonald's, Disney, and Coca-Cola corporations as agents of globalism or cultural imperialism, a new form of economic and political domination. Critics of globalism argue that any business enterprise capable of manipulating personal tastes will thrive, whereas state authorities everywhere will lose control over the distribution of goods and services. According to this view of world power, military force is perceived as hopelessly out of step or even powerless; the control of culture (and its production) is seen as far more important than the control of political and geographic borders. Certainly, it is true that national boundaries are increasingly permeable and any effort by nations to exclude global pop culture usually makes the banned objects all the more irresistible.

The commodities involved in the exchange of popular culture are related to lifestyle, especially as experienced by young people: pop music, film, video, comics, fashion, fast foods, beverages, home decorations, entertainment systems, and exercise equipment. Millions of people obtain the unobtainable by using the Internet to breach computer security systems and import barriers. "Information wants to be free" was the clarion call of software designers and aficionados of the World Wide Web in the 1990s. This code of ethics takes its most creative form in societies where governments try hardest to control the flow of information (*e.g.,* China and Iran). In 1999, when Serbian officials shut down the operations of Radio B92, the independent station continued its coverage of events in the former republic of Yugoslavia by moving its broadcasts to the Internet.

Questioning the future role of governments

The idea of a borderless world is reflected in theories of the "virtual state," a new system of world politics that is said to reflect the essential chaos of 21st-century capitalism. In *Out of Control* (1994), Kevin Kelly predicted that the Internet would gradually erode the power of governments to control citizens; advances in digital technology would instead allow people to follow their own interests and form trans-state coalitions. Similarly, Richard Rosecrance, in *The Rise of the Virtual State* (1999), wrote that military conflicts would be superseded by the flow of information, capital, technology, and manpower among states. Taking a countervailing approach, Martin Wolf wrote, in the January 2001 issue of *Foreign Affairs,* that the state was unlikely to disappear and could continue to be an effective basis of governance.

Arguments regarding the erosion of state sovereignty are particularly unsettling for nations that have become consumers rather than producers of digital technology. Post-Soviet Russia, post-Maoist China, and post-Gaullist France are but three examples of Cold War giants that face uncertain futures in the emerging global system. French intellectuals and politicians have seized upon anti-globalism as an organizing ideology in the absence of other unifying themes. In *Les cartes de la France à l'heure de la mondialisation* (2000; "France's Assets in the Era of Globalization"), French Foreign Minister Hubert Vedrine denounced the United States as a "hyperpower" that promotes "uniformity" and "unilateralism." Speaking for the French intelligentsia, he argued that France should take the lead in building a "multipolar world."

Globalization as a threat to national identity

Ordinary French citizens also were concerned about losing their national identity, particularly as the regulatory power of the European Union began to affect everyday life. Sixty percent of respondents in a 1999 *L'Expansion* poll agreed that globalization represented the greatest threat to the French way of life. Food, especially haute cuisine, is commonly regarded as the core of French culture. Yet, by the end of 2000 there were 857 McDonald's restaurants in France, collectively employing more than 30,000 people. The Big Mac may be a reviled symbol of cultural imperialism for French intellectuals, but the steady growth of fast food chains demonstrates that anti-globalism attitudes do not always affect economic behaviour, even in societies (such as France) where such sentiments are nearly universal. Like their counterparts in the United States, French workers are increasingly pressed for time. Fast food is convenient. The two-and-a-half-hour lunch is a thing of the past for most Parisians.

Anti-globalism movements and the Internet. Anti-globalism organizers are found throughout the world, not least in many management organizations. They are often among the world's most creative and sophisticated users of Internet technology. This is doubly ironic, for NGOs, as stated earlier, exhibit many of the characteristics of a global, transnational subculture; the Internet, moreover, is one of the principal tools that makes globalization feasible and organized protests against it possible. For example, Greenpeace, an environmentalist NGO, has orchestrated worldwide protests against genetically modified (GM) foods. Highly organized demonstrations appeared, seemingly overnight, in many parts of the world, denouncing GM products as "Frankenfoods" that pose unknown (and undocumented) dangers to people and to the environment. The bioengineering industry, supported by various scientific organizations, launched its own Internet-based counterattack, but the response was too late and too disorganized to outflank Greenpeace and its NGO allies. Sensational media coverage had already turned consumer sentiment against GM foods before the scientific community even entered the debate.

The anti-GM food movement demonstrates the immense power of the Internet to mobilize political protests. This power derives from the ability of a few determined activists



Demonstrators marching to protest against genetically modified organisms (GMOs), Montreal, January 2000.
Shaun Best—Reuters NewMedia, Inc./Corbis

to communicate with thousands (indeed millions) of potential allies in an instant. The Internet's power as an organizing tool became evident during the World Trade Organization (WTO) protests in Seattle, Wash., in 1999, in which thousands of activists converged on the city, disrupting the WTO meetings and drawing the world's attention to criticisms of global trade practices. The Seattle protests set the stage for similar types of activism in succeeding years.

THE ILLUSION OF GLOBAL CULTURE

Localized responses. For hundreds of millions of urban people the experience of everyday life has become increasingly standardized since the 1960s. Household appliances, utilities, and transportation facilities are increasingly universal. Technological "marvels" that North Americans and Europeans take for granted have had even more profound effects on the quality of life for billions of people in the less developed world. Everyday life is changed by the availability of cold beverages, hot water, frozen fish, screened windows, bottled cooking-gas, or the refrigerator.

Internal meanings

It is difficult to argue, however, that the globalization of technologies is making the world everywhere the same. The "sameness" hypothesis is only sustainable if one ignores the internal meanings that people assign to cultural innovations.

Borrowing and "translating" popular culture. The domain of popular music illustrates how difficult it is to unravel cultural systems in the contemporary world: Is rock music a universal language? Do reggae and ska have the same meaning to young people everywhere? American-inspired hip-hop (rap) swept through Brazil, Britain, France, China, and Japan in the 1990s. Yet Japanese rappers developed their own, localized versions of this art form. Much of the music of hip-hop, grounded in urban African American experience, is defiantly antiestablishment, but the Japanese lyric content is decidedly mild, celebrating youthful solidarity and exuberance. Similar "translations" between form and content have occurred in the pop music of Indonesia, Mexico, and Korea. It is also obvious to any casual listener of U.S. radio that Brazilian, South African, Indian, and Cuban forms have had profound effects on the contemporary American pop scene. The flow of popular culture is rarely, if ever, unidirectional.

Subjectivity of meaning—the case of *Titanic*. A cultural phenomenon does not convey the same meaning everywhere. In 1998, the drama and special effects of the American movie *Titanic* created a sensation among Chinese fans. Scores of middle-aged Chinese returned to the theatres over and over—crying their way through the film. Enterprising hawkers began selling packages of facial tissue outside Shanghai theatres. The theme song of *Titanic* became a best-selling CD in China. Chinese consumers purchased more than 25 million pirated (and 300,000 legitimate) video copies of the film.

One might ask why middle-aged Chinese moviegoers found themselves so emotionally involved with the story told in *Titanic*. Interviews among older residents of Shanghai revealed that many people had projected their own, long-suppressed experiences of lost youth onto the film. From 1966 to 1976 the Cultural Revolution convulsed China, destroying any possibility of educational or career advancement for millions of people. At that time, Communist authorities had also discouraged romantic love and promoted politically correct marriages based on class background and revolutionary commitment. Improbable as it might seem to Western observers, the story of lost love on a sinking cruise ship hit a responsive chord among the veterans of the Cultural Revolution. Their passionate, emotional response had virtually nothing to do with the Western cultural system that framed the film. In China, then, *Titanic* served as a socially acceptable vehicle for the public expression of regret by a generation of aging revolutionaries who had devoted their lives to building a form of socialism that had long since disappeared.

Chinese President Jiang Zemin invited the entire Politburo of the Chinese Communist Party to a private screening of *Titanic*, cautioning that *Titanic* could be seen as a Trojan horse, carrying within it the seeds of American cultural imperialism. Chinese authorities are not alone in their mistrust of Hollywood. There are those who suggest, as did Jiang, that exposure to films such as *Titanic* will cause people everywhere to become more like Americans. Yet anthropologists are wary of such suggestions, emphasizing instead the particular ways in which consumers use popular entertainment. The process of globalization looks far from hegemonic when one focuses on ordinary viewers and their efforts to make sense of what they see.

Interpretation of popular entertainment

Another case in point is anthropologist Daniel Miller's study of television viewing in Trinidad, which demonstrated that viewers are not passive observers. In 1988, 70 percent of Trinidadians who had access to a television watched daily episodes of *The Young and the Restless*, a series that emphasized family problems, sexual intrigue, and gossip. Miller discovered that Trinidadians had no trouble relating to the personal dramas portrayed in American soap operas, even though the lifestyles and material circumstances differed radically from life in Trinidad. Local people actively reinterpreted the episodes to fit their own experience, seeing the televised dramas as commentaries on contemporary life in Trinidad. The portrayal of American material culture, notably women's fashions, was a secondary attraction. In other words, it is a mistake to treat television viewers as passive.

THE TIES THAT STILL BIND

Local culture remains a powerful influence in daily life. People are tied to places, and those places continue to shape particular norms and values. The fact that residents of Moscow, Beijing, and New Delhi occasionally eat at McDonald's, watch Hollywood films, and wear Nike athletic shoes (or copies of them) does not make them "global." Outward appearances do not reveal the internal meanings that people assign to a cultural innovation. True, the standardization of everyday life will likely accelerate as digital technology comes to approximate the toaster in "user-friendliness." But technological breakthroughs are not enough to create a world culture. Studies in globalization show that people everywhere have an unquenchable desire to partake of the fruits of globalization while celebrating the inherent distinctiveness of their own cultures.

(Jm.L.W.)

BIBLIOGRAPHY. Texts providing theories of globalization include PETER L. BERGER, "Four Faces of Global Culture," *The National Interest*, no. 49, pp. 23–29 (Fall 1997); ARJUN APPADURAI, *Modernity at Large: Cultural Dimensions of Globalization* (1996); ULF HANNERZ, "Cosmopolitans and Locals in World Culture," in MIKE FEATHERSTONE (ed.), *Global Culture: Nationalism, Globalization, and Modernity* (1990), pp. 237–251; JOHN TOMLINSON, *Cultural Imperialism: A Critical Introduction* (1991); and MALCOLM WATERS, *Globalization*, 2nd ed. (2001). Discussions of food and global dietary trends are found in JAMES L. WATSON (ed.), *Golden Arches East: McDonald's in East Asia* (1997); SUSAN TAX FREEMAN, "The Capsicums in Old World Culinary Structures," in LEONARD PLOTNICOV and RICHARD SCAGLION (eds.), *Consequences of Cultivar Diffusion* (1999), pp. 75–83; and SUCHETA MAZUMDAR, "The Impact of New World Food Crops on the Diet and Economy of China and India, 1600–1900," chapter 3 in RAYMOND GREW (ed.), *Food in Global History* (1999), pp. 58–78. Studies of television are discussed in ROBERT C. ALLEN (ed.), *To Be Continued—Soap Operas Around the World* (1995); and DANIEL MILLER, "The Young and the Restless in Trinidad: A Case of the Local and the Global in Mass Consumption," chapter 10 in ROGER SILVERSTONE and ERIC HIRSCH (eds.), *Consuming Technologies: Media and Information in Domestic Spaces*, updated ed. (1994), pp. 163–182. Broader issues of popular culture are treated in WALTER LAFEBER, *Michael Jordan and the New Global Capitalism* (1999); TIMOTHY D. TAYLOR, *Global Pop: World Music, World Markets* (1997); ERIC SMOODIN (ed.), *Disney Discourse: Producing the Magic Kingdom* (1994); AVIAD E. RAZ, *Riding the Black Ship: Japan and Tokyo Disneyland* (1999); and IAN CONDRY, "The Social Production of Difference: Imitation and Authenticity in Japanese Rap Music," chapter 8 in HEIDE FEHRENBACH and UTA G. POIGER (eds.), *Transactions, Transgressions, Transformations: American Culture in Western Europe and Japan* (2000), pp. 166–186.

Goethe

Poet, novelist, playwright, and natural philosopher, Goethe is the greatest figure of the German Romantic period and of German literature as a whole. He was perhaps the last European to achieve the many-sidedness of the great Renaissance personalities: critic, journalist, painter, theatre manager, statesman, educationist, natural philosopher. The bulk and diversity of his output are phenomenal: his writings on science alone fill about 14 volumes.

By courtesy of the Bayerische Staatsgemäldesammlungen, Munich. Neue Pinakothek



Goethe, oil painting by Joseph Karl Stieler, 1828. In the Neue Pinakothek, Munich.

Early life and influences. Johann Wolfgang von Goethe was born on Aug. 28, 1749, in Frankfurt am Main. He came of middle-class stock, the *Bürgertum* that he praised as a breeding ground of the finest culture. His father, a retired lawyer, was able to lead a life of cultured leisure. Goethe's mother was the daughter of a *Bürgermeister* (mayor) of Frankfurt. Of eight children, only Wolfgang, the firstborn, and his sister, Cornelia, survived. In his autobiography, *Dichtung und Wahrheit* (1811–32; *Poetry and Truth from My Own Life*), Goethe left an unforgettable picture of a happy childhood.

In October 1765 Goethe was sent to study law at his father's alma mater, the University of Leipzig, where a world of elegance and fashion daunted the young provincial. The Frenchifying influence of the critic J.C. Gottsched still dominated the theatre and provided a repertory of the best plays of contemporary Europe. But C.F. Gellert, poet and author of fables and hymns, presented the new sensibility of Edward Young, Laurence Sterne, and Samuel Richardson. Gellert's literary influence was reinforced by the robust elegance and ironic sagacity of the works of C.M. Wieland.

The literary harvest of Goethe's Leipzig period manifested itself in a songbook written in the prevailing Rococo mode—songs praising love and wine in the manner of the Greek poet Anacreon. Appropriately titled *Das Leipziger Liederbuch* (1770; *The Leipzig Song Book*), it was ostensibly but not passionately inspired by the daughter of the wine merchant at whose tavern he took his midday meal. The same note is struck in two verse plays, *Die Laune des Verliebten* (1767, printed 1806; "The Mood of the Beloved") and a more sombre farce, *Die Mitschuldigen* (1768, printed 1787; "The Accomplices"), which foreshadows the psychological preoccupations of later works. From then on, the Rococo was to reappear in the setting of *Torquato Tasso* (1790) and *Die Wahlverwandtschaften* (1809; *Elective Affinities*); he paid tribute to its charm in *Anakreons Grab* (1806; "Anacreon's Grave") and amalgamated it with Eastern influence in the enchanting poems of his *West-östlicher Divan* (1819; *Poems of the West and the East*).

Works of the storm and stress period. Goethe's stay in Leipzig was cut short in 1768 by severe illness. A long convalescence at home fostered introspection and religious mysticism. On his recovery it was decided that he should pursue legal studies in Strassburg. In this German capital of a French province, he experienced a reaction against the cosmopolitan atmosphere of Leipzig and under the impact of the great cathedral proclaimed his conversion to the Gothic German ideal. More decisive still was the influence of J.G. Herder, who spent the winter of 1770–71 there undergoing treatment for his eyes. From him Goethe learned a new view of the artist as a creator fashioning forms expressive of feeling; a new theory of poetry as the original and most vital human language; the virtues of a new style, that of the *Volkslied* (folk song) and the poetry of "primitive" peoples as enshrined in the Bible, the epics of Homer, and the poems attributed (falsely) to Ossian, a 3rd-century Celtic poet.

In writing the *Geschichte Gottfriedens von Berlichingen mit der eisernen Hand dramatisiert* (1771; "The Dramatized History of Gottfried von Berlichingen of the Iron Hand"), Goethe was deliberately vying with Shakespeare. With the publication in 1773 of *Götz von Berlichingen*, a radically tautened version of that "History," the Shakespeare cult was launched, and the Sturm und Drang (storm and stress) movement was provided with its first major work of genius. The manifesto of the movement—"Von deutscher Art und Kunst" ("Concerning German Nature and Art")—was heralded by Goethe's enthusiastic *Rede zum Schakespears Tag* ("Conversation from Shakespeare's Day"). It appeared after Goethe's return to Frankfurt in August 1771.

The Sturm und Drang movement

Literature soon won the day over law, and Goethe's impassioned yet self-ironic ode in free verse, "Wandrers Sturmlied" ("Wanderer's Storm Song"), is testimony both to a recently inspired admiration for Pindar and to a hesitant certainty that he himself might be destined for greatness. And he experienced a new passion, Charlotte Buff, who was engaged to be married. She was enshrined in *Die Leiden des jungen Werthers* (1774; *The Sorrows of Young Werther*). But the real theme of *Werther* is what the 18th century called Enthusiasm: the fatal effects of a predilection for absolutes. The title has been trivialized in translation: *Sorrows* (instead of "Sufferings"), which obscures the allusion to the Passion of Christ and individualizes what Goethe himself thought of as a "general confession," in a tradition going back to St. Augustine.

Besides *Werther* and *Götz*, the period 1771–75 saw the appearance of a number of magnificent hymns, including "Der Wandrer" ("The Wanderer"); the inception of *Egmont* (1788) and *Faust* (this so-called *Urfaust*, or "original" version of *Faust*, was discovered by a lucky chance in 1887); the completion of *Clavigo* (1774), a play of more "regular" form on a theme of the French playwright Beaumarchais, and of *Stella* (1775), with its conciliatory ending of a *mariage à trois*, subsequently conventionalized into tragedy. Two operettas, *Erwin und Elmire* (1775) and *Claudine von Villa Bella* (1776), reflect a return to the elegance of Rococo inspired by Goethe's short-lived betrothal to Lili Schönemann, daughter of a banker.

The mature years at Weimar. In 1775 Goethe went to Weimar on a visit to the reigning duke, the major turning point of his life. Weimar remained his home—despite Napoleon's invitation to Paris—until his death there on March 22, 1832. From now on, mastery of life became his chief concern; and *Wilhelm Meisters Lehrjahre* (1795–96; *Wilhelm Meister's Apprenticeship*), the title he eventually gave his next novel, suggests the long apprenticeship such

mastery involves. (Thomas Carlyle, the Scottish historian and essayist, did a distinguished translation of the novel.)

Goethe developed a passionate devotion to the wife of a court official, Charlotte von Stein. The sublimation she increasingly enforced on him, though irksome, could inspire works ranging from the almost psychoanalytical probings of "Warum gabst du uns die tiefen Blicke?" ("Why did you give us the deep glances?") to such well-loved lyrics as "An den Mond" ("To the Moon") and the two exquisite "Wanderers Nachtlieder" ("Wanderer's Night Songs"). In these and other poems of this period, nature has ceased to be a mere reflection of man's moods and has become something existing in its own right. This new "objectivity" is in tune with Goethe's growing scientific preoccupations. Yet such is his versatility that he could, when he chose, compose ballads such as "Erlkönig" ("Elf King"), in which nature bears the projection of unconscious forces.

Journey to Italy

In September 1786, in dramatic secrecy, Goethe set out on a long-postponed Italian journey—recounted in *Italienische Reise* (1816–17), which was translated in the 20th century as *Italian Journey 1786–88* by the poet W.H. Auden and Elizabeth Mayer. Goethe sought among other things a key to the world of Homer, which he recaptured in a glorious dramatic fragment, *Nausikaa* (1787). He sought and found the *Urmensch*, or archetypal man, in the forms of Greek antiquity, and in these landscapes there also came to his mind the extension of this idea to plants as well. In his literary work these pursuits led to the creation of beings who are individual manifestations but of a clearly discernible type; to themes that are universal and timeless but treated in a highly differentiated way; to the measured cadences of verse that are yet vibrant with personal passion.

This new conception of form is apparent in the revision of the four plays he had taken with him to Italy. *Faust, Ein Fragment* ("Faust, a Fragment"), published in 1790, is quite clearly a step in the direction of the stupendous cultural symbol the play would eventually become. *Egmont,* though not actually cast into verse, is raised to the level of poetic drama by a thickening of the verbal texture, so that when music finally takes over it seems the inevitable culmination of a gradual convergence and sudden contraction of themes.

In *Torquato Tasso* such linguistic density is carried to lengths possible only in verse. The tragic conflict in this play about a poet arises from misunderstandings about the various modes of language, and the temperamental clashes are presented as concomitants of this. By placing the poet in a society that, far from being indifferent or hostile, cherishes him and values his work, Goethe has thrown into sharpest relief the incurable "discrepancy" between poet and world.

But it was perhaps *Iphigenie auf Tauris* (1787) that benefitted most from Goethe's encounter with classical antiquity, though Schiller was right in calling it "astonishingly modern and un-Greek." Like *Tasso,* it treats of the problems of communication. But it also concerns man's power to free himself from his myths by recognizing them as projections of his own unconscious, his power to break the chain of events that seems to determine his present (symbolized in the monotonously regular crime sequence of the race of Tantalus) by a reorientation of outlook. In its synthesis of Greek and Christian values, its elevation of the physical to the spiritual through the identification of Iphigenie with the divine sister, Diana, this play represents the highest achievement of 18th-century humanism.

Synthesis of Greek and Christian values

The chief lyrical product of the Italian journey was the *Römische Elegien* (written 1788–89, published 1795; *Roman Elegies and Other Poems*). The inspiration for these elegies was Christiane Vulpius, daughter of a humble official, with whom Goethe had begun living soon after his return from Italy, although he did not marry her until 1806. Christiane bore him several children.

A return visit to Italy in 1790 brought nothing but disappointment, and a restlessness aggravated by the revolutionary events in the outer world. The *Venezianische Epigramme* (1790; published in *Roman Elegies and Venetian Epigrams*) reflect something of this discontent. In 1792 Goethe accompanied his duke on the disastrous campaign into France and wrote up his experiences in two still very readable war books, *Campagne in Frankreich 1792* and *Die Belagerung von Mainz 1793* (both 1822; published in *Miscellaneous Travels of J.W. Goethe*). The French Revolution forms the background of his Homeric treatment of the refugee problem, *Hermann und Dorothea* (1797), and it fills the whole canvas of *Die Natürliche Tochter* (1804; "The Natural Daughter").

Schiller and the Classical ideal. The human and spiritual isolation in which Goethe found himself on his return from Italy was unexpectedly relieved by the development of a friendship with Friedrich von Schiller. Some of the works Goethe produced during the next few years are embodiments of the Classical ideal espoused by Schiller (and by Goethe). *Hermann und Dorothea,* one of the best loved, is Goethe's attempt to "produce a Greece from within." The characters are types—except for the hero and heroine, they have no proper names, and even the hero's and heroine's are symbolic—and like those of the *Odyssey* they vindicate peace and home and the domestic virtues. Yet, as always in Goethe's works, these are shown as never secure for long, as constantly in need of being fostered by the efforts to be human and humane. In the Helena act of *Faust, Part II* (1832), in which the meeting and mating of Faust and Helen of Troy marks the synthesis of paganism and Christianity, of Greece and Germany, he captured the Greek spirit so successfully that competent critics hold that if translated into Attic Greek it might well pass for a lost fragment of the Athenian stage. A never completed epic, *Achilleis* (1808), is his last attempt to "be a Greek after his own fashion."

Friendship with Schiller

Other works of this period are in tune with Schiller's growing conviction that the only future for literature in a world that increasingly clamoured for the naturalistic and the tendentious lay in a hermetic closing of the poetic world by a frank introduction of symbolic devices. *Wilhelm Meisters Theatralische Sendung* (*Wilhelm Meister's Theatrical Mission;* a manuscript of this version turned up in 1910) is now widened to a vocation for life and is wholly in tune with Goethe and Schiller's conviction that art, though not the handmaid of either truth or morality, has nevertheless its own peculiar part to play in making better men and better citizens. (Selections from their correspondence can be found in *Correspondence between Goethe and Schiller,* 1994.)

It was Schiller, too, who turned Goethe's thoughts to the continuation of *Faust* and discerned the difficulties involved in reconciling this "barbarous composition" with their Classical ideal, in blending the evident seriousness of its "idea" with that element of "play" that was the prerequisite of the art of the future. By his insistence on such problems, he inspired the fictional framework of *Faust*'s "Prelude on the Stage" no less than the philosophical framework of the "Prologue in Heaven."

Goethe's relation to the Romantics. With Schiller's death in 1805, Goethe felt he had lost "the half of his existence," and he wrote a magnificent tribute to his great friend in *Epilog zu Schillers Glocke* (1805; "Epilogue to Schiller's Bells"). His intellectual loneliness was eased by his relations to the new school of Romantics then flourishing in Jena, for they had much in common. In *Die Wahlverwandtschaften* he drew heavily for his thematic material upon their preoccupation with "the night side of nature."

The Romantic movement

By their translations the Romantics were opening up the literary treasures of the world, and *Weltliteratur* was to become one of Goethe's most treasured concepts. Its aim was, as he put it, to advance civilization by encouraging mutual understanding and respect whether through translation or criticism (his own attempts to interpret Serbian poetry to the Germans is an excellent example of the latter) or through the blending of different literary traditions. Two great ballads, "Der Gott und die Bajadere" ("The God and the Dancing Girl") and "Paria" ("Outcast"), and two exquisite cycles, the late and lesser known *Chinesisch-Deutsche Jahres- und Tageszeiten* (1830; "Chinese-German Hours and Seasons") and the *West-östlicher Divan,* are his own outstanding attempts to marry East with West.

The last decade. In his last years, Goethe found himself a world figure, and little Weimar drew a constant stream

of pilgrims. Goethe's continued openness to the world around him and its possibilities is nowhere more apparent than in *Wilhelm Meisters Wanderjahre* (1821–29; *Wilhelm Meister's Years of Travel*), with its commitment to social and technological progress, to a type of education better adapted to modern specialization than the old humanistic studies, to a world no longer centred wholly in Europe—a major "complication" of his plot is a resettlement plan for emigrants in the land of the future ("Amerika, du hast es besser!" ["America, you are better off!"]).

Faust

Faust, like *Wilhelm Meisters Wanderjahre,* is often decried as formless. The array of lyric, epic, dramatic, operatic, and balletic elements, of almost every known metre, of styles ranging from Greek tragedy through medieval mystery, Baroque allegory, Renaissance masque, commedia dell'arte, and the "temerities of the English stage," to something akin to the modern revue, all suggest a deliberate attempt to make these various forms a vehicle of cultural comment rather than any failure to create a coherent form. And the content with which Goethe invests his forms bears this out. He draws on an immense variety of cultural material—theological, mythological, philosophical, political, economic, scientific, aesthetic, musical, literary—for the more realistic Part I no less than for the more symbolic Part II (first published posthumously in 1832): if Faust's wooing of Helena in the "Classic-Romantic Phantasmagoria" (as the first publication of the scene in 1827 called it) is accomplished by teaching her the unfamiliar delights of rhymed verse, his seduction of Gretchen is firmly set in the long tradition of erotic mysticism going back to the Song of Solomon. The Faust myth is here made the medium of a profoundly serious but highly ironic commentary on our cultural heritage, presented as a drama of the diverse potentialities that coexist in Western civilization. This Faust, unlike his creator, is the very type of Western man, with two souls warring within his breast and a restlessly inquiring spirit. If the seal of approval is set on a spirit that has eluded Mephisto's every effort to lull him into sloth, the evil into which it led him is not condoned. Indeed, none of Goethe's conciliatory endings, except that of *Iphigenie,* really removes the sting of tragedy.

Similarly, Goethe was profoundly aware of the dual nature of music and as suspicious as Plato of its orgiastic power. As in every art he looked for the taming of the Dionysian by the Apollonian, nowhere more movingly symbolized than by the taming of the lion through the piping of the little child in his *Novelle* of 1828. Increasingly he turned to music for assuagement of his own suffering. His *Trilogie der Leidenschaft* (1823–27; "Trilogy of Passion") is at once the lyrical precipitate of an old man's anguished love for a girl of 18 and a tribute to the cathartic effect of this "heavenly art," which restores to life even as it soothes. His *Zauberflöte, Zweiter Teil* (1794) is a tribute to his favourite Mozart's *Magic Flute:* Mozart would, he thought, have been the ideal composer for *Faust.*

Scientific interests

By common consent, *Faust* is one of the supreme, if as yet unclassified, achievements of literature. But there were moments when Goethe rated his scientific work higher than all his poetry. The usefulness of the Psycho-Physiological Section of his *Farbenlehre* (1805–10; *Goethe's Colour Theory*), together with his study *Entoptische Farben* ("Entoptic Colours"), is generally acknowledged, while the Historical Section is something of a pioneer work in the writing of the history of science. His work in botany and biology is less controversial. His *Metamorphose der Pflanzen* (1790; *The Metamorphosis of Plants*) is a model of presentation, and the drawings in it are a botanist's delight. His main thesis, that all the parts of the plant are modifications of a type-leaf, has met with a measure of acceptance, though his categorical neglect of the root is regarded as an unscientific exclusion of a possible area of relevance. His discovery in 1784, arrived at independently even if he was not the first to make it, of a recognizable os intermaxillare (the premaxilla of modern anatomists) in the human species was yet another result of his sustained quest for unity and continuity in nature and caused Charles Darwin to hail him as a forerunner.

Scrupulous awareness of his own mental operations was of paramount importance in morphology, the science Goethe founded and named. Morphology, as he understood it, was the systematic study of formation and transformation—whether of rocks, clouds, colours, plants, animals, or the cultural phenomena of human society—as these present themselves to sentient experience. Goethe was aiming at an understanding of nature in all its qualitative manifestations; and one of his most impassioned pleas is for a concert of all the sciences, a cooperation of all types of method and mind.

This impulse, to find a scientific as well as an aesthetic corrective to the inevitably esoteric tendencies of specialization, is nowhere more apparent than in his two elegies on plant and animal metamorphosis in which he tries to present to imagination and feeling what has been understood by the mind. They eventually took their place in a cycle of philosophical poems entitled *Gott und Welt* ("God and World"). Goethe was and remained a grateful heir of the Christian tradition—*bibelfest,* rooted in the Bible. From this centre he extended sympathetic understanding to all other religions. "Panentheism" has been proposed as a more exact term for his belief in a divinity at once immanent and transcendent.

Truth as the embracing of opposites

Assessment. Goethe's Werther knew that the realities of existence are rarely to be grasped by Either-Or. And the reality of Goethe himself certainly eludes any such attempt. He was, as befits a son of the Enlightenment, wholly committed to the adventure of science; but he stood in awe and reverence before the mystery of the universe. Goethe nowhere formulated a system of thought. He was as impatient of the sterilities of logic chopping as of the inflations of metaphysics, though he acknowledged his indebtedness to many philosophers, including Immanuel Kant. But here again he was not to be confined. Truth for him lay not in compromise but in the embracing of opposites. And this is expressed in the form of his *Maximen* ("maxims"), which, together with his *Gespräche* ("conversations"), contain the sum of his wisdom. As with proverbs, one can always find among them a twin that expresses the complementary opposite. And they have something of the banality of proverbs, too. But it is, as André Gide observed, "*une banalité supérieure.*" What makes it "superior" is that the thought has been felt and lived and that its formulation betrays this. Not an ascetic, a mystic, a saint, or a recluse, not a Don Juan or a poet's poet but a man who to the best of his ability had tried to achieve the highest form of *l'homme moyen sensuel*—which is perhaps what Napoleon sensed when after their meeting in Erfurt he uttered his famous "*Voilà un homme!*"

BIBLIOGRAPHY. English translations of representative works from his oeuvre are in *Goethe's Collected Works,* 12 vol. (1983–89, reissued 1994–95), published by Suhrkamp; and *Goethe's Works,* 14 vol. (1848–90), in Bohn's Standard Library. Bibliographic details are available in EUGEN OSWALD, *Goethe in England and America,* 2nd ed., rev. and enlarged by LINA OSWALD and ELLA OSWALD (1909).

The best introduction in English to Goethe's life is GEORGE H. LEWES, *The Life and Works of Goethe, with Sketches of His Age and Contemporaries from Published and Unpublished Sources,* 2 vol. (1855; also published as *Life of Goethe*), available also in later editions and printings. Also of interest are NICHOLAS BOYLE, *Goethe: The Poet and the Age* (1991–); BENEDETTO CROCE, *Goethe,* trans. by EMILY ANDERSON (1923, reprinted 1973; originally published in Italian, 1919); and GEORGE BRANDES, *Wolfgang Goethe,* trans. by ALLEN W. PORTERFIELD, 2 vol. (1924, reissued 2 vol. in 1, 1936; originally published in Danish, 1915). Critical analyses of his works are found in THOMAS MANN, *Three Essays,* trans. from German by H.T. LOWE-PORTER (1929, reissued 1932); ELIZABETH M. WILKINSON and LEONARD A. WILLOUGHBY, *Goethe: Poet and Thinker* (1962); and RONALD PEACOCK, *Goethe's Major Plays* (1959, reprinted 1966). Other studies include GEORG LUKÁCS (GYÖRGY LUKÁCS), *Goethe and His Age* (1968, reprinted 1979; originally published in German, 1947).

(E.M.Wn./Ed.)

The Forms of Government: Their Historical Development

Most of the key words commonly used to describe governments, words such as monarchy, oligarchy, and democracy, are of Greek or Roman origin. They have been current for more than 2,000 years and have not yet exhausted their usefulness. This suggests that mankind has not changed very much since they were coined; but such verbal and psychological uniformity must not be allowed to hide the enormous changes in society and politics that have occurred. The earliest analytical use of the term monarchy occurred in ancient Athens, chiefly in Plato's dialogues, but even in Plato's time the word was not self-explanatory. There was a king in Macedon and a king in Persia, but the two societies, and therefore their institutions, were radically different. To give real meaning to the word monarchy in these two instances, it would be necessary to investigate their actual political and historical contexts. Any general account of monarchy required then, and requires today, an inquiry as to what circumstances have predisposed societies to adopt monarchy, and what have led them to reject it. So it is with all political terms.

This article is divided into the following sections:

Primitive government 141
  Agricultural society
  The spread of civilization
Greece 142
  The city-state
  Monarchy, oligarchy, democracy
Rome 142
  The republic
  The empire
The Middle Ages 143
  Dissolution and instability
  Feudalism
  The rise of law and the nation-state
Emergence of the modern world 144
  The failure of absolutism
  Representation and constitutional monarchy
  The American and French republics
  Nationalism and imperialism
20th-century models 146
  The Soviet state
  Liberal democracy
Bibliography 147

PRIMITIVE GOVERNMENT

Agricultural society. So long as humankind were few, there was hardly any government. The division of function between ruler and ruled occurred only, if at all, within the family. The largest social groups, whether tribes or villages, were little more than loose associations of families, in which every elder or family head had an equal voice. Chieftains, if any, had strictly limited powers; some tribes, no doubt, did without chieftains altogether. This prepolitical form of social organization may still be found in undeveloped regions of the world, such as the Amazonian jungle in South America or the Upper Nile Valley in Africa.

Pre-political societies

The rise of agriculture began to change this state of affairs. In the land of Sumer (modern Iraq) the invention of irrigation necessitated grander arrangements. Control of the flow of water down the Tigris and Euphrates rivers had to be coordinated by a central authority, so that downstream fields could be watered as well as those further up. It became necessary also to devise a calendar, so as to know when the spring floods might be expected. As these skills evolved, society evolved with them. In early Sumer, it is reasonable to assume, the heads of the first cities, which were little more than enlarged villages, only gradually assumed the special attributes of monarchy, the rule of one; and the village council only gradually undertook a division of labour, so that some specialized as priests and others as warriors, farmers, or tax gatherers (key figures in every civilized society). As organization grew more complex, so did religion: an elaborate system of worship was necessary to propitiate the quite elaborate family of gods who, it was hoped, would protect the city from attack, from natural disaster, and from any questioning of the political arrangements deemed necessary by the ruler group.

Unfortunately, but inevitably, the young cities of Sumer quarrelled over the distribution of the rivers' water, and their wealth excited the greed of nomads outside the still comparatively small area of civilization. War, perhaps the most potent of all forces of historical change, announced its arrival, and military leadership became at least as important an element of kingship as divine sanction. It was to remain so throughout the long history of monarchy: whenever kings have neglected their military duties they have endangered their thrones. The wars of Sumer also laid bare another imperative of monarchy—the drive for empire, arising from the need to defend and define frontiers by extending them, and the need to find the means to pay for troops and weapons, whether by the plunder of an enemy or by the conquest of new lands.

Military and imperial imperatives

The spread of civilization. The history of Old World monarchy, and indeed of civilization, was to consist largely of variations on these patterns for four or five millennia. Trade contacts carried the principles of civilization to Egypt and to India (China seems to have evolved independently); and everywhere, once the social order was established, the problem of defending it became paramount. For although the broad zone of civilization spread steadily, so that by the reign of the Roman emperor Trajan (AD 98–117) there was a continuous band of civilized societies from Britain to the China Sea, it was always at risk from the barbarian nomads who roamed the great steppelands of central Eurasia. These nomads had retained the loose and simple institutions of humanity's infancy, but they had in other ways evolved as rapidly and successfully as the cities themselves (and partly under the cities' influence). The steppe was horse country, and, armed with bows and arrows, the barbarians of all epochs were marvelously swift and deadly light cavalry. They fought constantly among themselves for pasturage, and the losers were forever being driven west, south, and east, where they all too often overcame such defenses as the farms and cities of civilization could muster against them.

The nomads' military challenge was never sufficient to overturn civilization entirely. Either the invaders would overrun the settled lands and then adopt civilized customs, or the frontier defenses would prove strong enough to hold them off. There were even long periods of peace, when the barbarian threat was negligible. It was at such times that the spontaneous ingenuity of mankind had greatest play, in politics as in everything else. But it is noteworthy that, in the end, what may be described as the ancient norm always reasserted itself, whether in Europe, the Middle East, India, or China. Military crises—civil war or barbarian invasion or both—recurred and necessitated the strengthening of government. The effort to secure a measure of peace and prosperity required the assertion of authority over vast distances, the raising of large armies, and the gathering of taxes to pay for them. These requirements in turn fostered literacy and numeracy and the emergence of what later came to be called bureaucracy, government

by official. Bureaucratic imperialism emerged again and again and spread with civilization. Barbarian challenge occasionally brought it low but never for very long. When one city or people rose to hegemony over its neighbours, it simply incorporated their bureaucracy with its own. Sumer and Babylon were conquered by Assyria; Assyria was overthrown by the Medes, in alliance with Babylon and nomad Scythians; the empire of the Medes and Persians was overthrown by Alexander of Macedon; the Macedonian successor states were conquered by Rome, which was in due course supplanted in the Middle East and North Africa by the Islāmic Caliphate of Baghdad. Conquerors came and went; life for their subjects, whether peasants or townsmen, was not much altered by anything they did, so long as the battles occurred at safely remote sites.

Bureaucratic monarchy

Nevertheless, from time to time experiments were made, for no monarchy commanded the resources to rule all its subjects directly. So long as they paid tribute punctually, local rulers and local communities were perforce left very much to govern themselves. Even if they did not pay, the effort required to mount a military operation at a distance from the imperial centre was so great that only in exceptional circumstances would it be undertaken, and even then it might not succeed, as the kings of Persia found when they launched punitive expeditions against mainland Greece at the beginning of the 5th century BC. So, in normal times, the inhabitants of the borderlands had extensive freedom of action.

Although civilization, as its advantages became clear, spread west and northwest out of Asia, bureaucratic monarchy could not easily follow it, for the sea was becoming a historical factor as important as the steppe and the great irrigable rivers. Tyre and Sidon, maritime cities of Phoenicia, had long exploited their coastal situation, not only to remain independent of the landward empires but also to push across the sea, even beyond the Strait of Gibraltar, in quest of trade. Their daughter-cities, Carthage, Utica, and Cádiz, were the first colonies, but primitive communications made it impossible for Phoenicia to rule them from afar.

GREECE

The city-state. The Phoenician example was followed by the Greeks. The Greeks were originally Indo-European nomads, who gradually made their way down to the Aegean and there took to the sea. They built on the achievements of earlier peoples, even taking over the first bureaucratic monarchy to appear on European soil: the Minoan civilization of the island of Crete, which succumbed to the invaders about 1450 BC. Continuing invasions from the north overthrew the mainland kingdoms of Mycenae, Tiryns, and Pylos in about 1200 BC. The so-called dark ages of Greece that then began lasted until the 8th century BC, by which time the Greeks had not only recovered literacy, by adapting the Phoenician alphabet, and begun to found overseas colonies, but had also brought the city-state to something near maturity. This form of government was the great political invention of classical antiquity.

Mediterranean geography is such that every little fishing village had to be able to defend itself against attack from land or sea, for outside help could not reach it easily. A man's dependence on his community, for physical as well as economic survival, was therefore obvious and complete.

The *polis*

The city (*polis*) had first claim on his labour and loyalty, a claim that was usually freely recognized. It was this reality that led Aristotle (who himself came from just such a small commonwealth, Stageira, or Stagirus) to define man as a political animal.

Coastal mountain ranges made it difficult for any community to dominate more than a few square miles of land. They also for a long time deterred the rise of an empire to federate and control all the cities. If a few of these centres nevertheless rose to imperial greatness, like Tyre and Sidon before them, it was in the main because, like their Phoenician predecessors, they traded across the sea successfully. Athens, for example, exported olive oil, silver, and pottery. The profits of this trade enabled it to build a great navy and formidable city walls. Athenian ships defeated Persia (480 BC) and won a small empire in the Aegean Sea; the combination of ships and walls enabled Athens long to defy, and nearly to defeat, its chief rival among the Greek cities, Sparta. Even after Sparta's triumph at the end of the Peloponnesian War (404 BC), Athens remained an independent sovereign state until its defeat by Philip of Macedon at the battle of Chaeronea (338 BC). In short, during the period of its prime Athens was free to make what experiments it liked in the realm of government, and to that period are owed not just the first example of successful democracy in world history but also the first investigations in political thought.

Monarchy, oligarchy, democracy. No Athenian believed that he had anything to learn from the bureaucratic monarchies of the East, which were incompatible with Greek notions of citizenship. All monarchies, indeed, seemed bound to deteriorate into such tyrannies. Self-defense necessitated that every adult male be required, and indeed be willing, to fight when called on; in return he had to be given some measure of respect and of personal autonomy—in a word, freedom. To protect that freedom, government was necessary; anarchy had no attractions for any Greek. The central question of politics, then, was the distribution of power among the citizens. Was Greek freedom best preserved and defined by the rule of the few or by that of the many? On the whole the great names favoured aristocracy, the rule of the best. Plato believed that the object of politics was virtue, and that only a few would ever thoroughly understand the science by which virtue could be attained. These trained few, then, should rule. Aristotle, his pupil, seems to have put the cultivation of the intellect highest among human goods; and he believed—quite reasonably, given the limited resources then available—that this fruit of civilization could be reaped only among a leisure class supported by the labours of the many. In return for their leisure the gentry would agree to sacrifice some of their time to the tedious business of governing, which only they would be sufficiently disinterested and well informed to do successfully. Neither of these apologies for oligarchy had any success in practice. The democrats carried the day, at any rate in Athens and its allied cities. In return for playing their parts as soldiers or sailors, ordinary Athenians insisted on controlling the government.

Views of Plato and Aristotle

The result was impressive. The people were misled by demagogues; they were intolerant enough to put Plato's master, Socrates, to death; they were envious of all personal distinction; and of their three great wars (against Persia, Sparta, and Macedon), they lost two. Furthermore, passionate devotion to the idea that Athens was the greatest of all cities, the school of Greece, and the wonder of civilization, misled them into basing their society in large part on slave labour, into wanton imperial adventure abroad, and into denying Athenian citizenship to all who were not born to it (even Aristotle), however much they contributed to the city's greatness and however much more they might have done. The foundations of Athenian democracy were narrow and shallow and fragile. But to say all this is only to say that the city could not entirely shake off the traditions of its past. Its achievement was the more remarkable for that. Seldom since has civilized humanity equalled democratic Athens, and until the last the city was satisfactorily governed by law and by popular decision. It owed its fall less to any flaw than to the overwhelming force that was mounted against it.

For to the north of Hellas proper a new power arose. Greek civilization had slowly trained and tamed the wild men of Macedon. Their king Philip forged them into a formidable army; he and his son Alexander then seized the opportunity open to them. History and geography made it impossible for the Greek cities to unite, and so they hanged separately. It seemed as if the city-state had been but a transient expedient. Henceforward Athens and Sparta would take their orders from the Great King.

ROME

The republic. As it turned out, the city-state had barely begun to display its full political potential. To the west, two non-Greek cities, Carthage and Rome, began to strug-

gle for mastery, and after the defeat of Hannibal at Zama (202 BC) Rome emerged as the strongest state in the Mediterranean.

The Greeks did not know how to classify Rome. The Greek historian Polybius, who chronicled Rome's rise, suggested that its constitution was such a success because it was a judicious blend of monarchy, aristocracy, and democracy. The Romans, a conservative, practical people, showed what they thought of such abstractions by speaking only of an unanalyzed "public thing"—*res publica*—and thus gave a new word to politics. With this focus the patriotism of the city-state reached its greatest intensity. The Romans were deeply attached to their traditions, and these traditions all taught the same lesson. For example, the legendary hero Mucius Scaevola gave his right hand to the flames to prove that there was nothing a Roman would not endure for his city. This passionate devotion to Rome's survival was tested again and again in war. All the tales of early Rome turn on battle. With dour persistence the peasants who had gathered for self-protection on the seven hills resisted every invader, fought back after every defeat, learned from all their mistakes, and even, however reluctantly and belatedly, modified their political institutions to meet the new needs of the times as they arose. Polybius was right: power in Rome was indeed shared among the people, the aristocracy, and the consuls, the executive officers of the republic who had replaced the kings. The claims of the many and of the few were fought out at election time, when the world's first clearly identifiable political parties appeared. Until Rome's late decadence, the results of elections were universally respected, and the triumphant alliance of the few and the many against the world was proclaimed in the letters blazoned on the city's buildings and battle standards: SPQR, for Senatus Populusque Romanus, the Senate and the people of Rome.

The Roman *res publica*

Like Athenian democracy, this system worked well for a surprisingly long time, and if the chief Athenian legacy was the proof that politics could be understood and debated logically and that under the right conditions democracy could work, Rome proved that the political process of competition for office and public discussion of policy was a valuable thing in itself.

Nevertheless, the Roman republic had been forged in a grim world. Its power had gradually extended over Italy in wars, always supposedly of self-defense. It is not surprising that what impressed the world most about the city was its military strength rather than its political institutions, even though the two were intimately related. As the weakness of Rome's neighbours became apparent, the Romans began to believe in their mission to rule, "to spare the conquered and war down the proud," as the poet Virgil put it. Military strength, in short, led to military adventurism. By the 1st century BC, Rome, having become a naval power as well as a military one, had conquered the whole Mediterranean basin and much of its hinterland. The strains of empire building made themselves felt. The Roman armies, no longer composed of citizens temporarily absent from the plow or the workshop but of lifetime professionals, were now loyal to their generals rather than to the state; and these generals brought on civil war as they competed to turn their foreign conquests into power at home. The population of Rome swelled, but economic growth could not keep pace, so that many citizens became paupers dependent on the public dole. The aristocrats appointed to govern the provinces saw their postings merely as opportunities to get rich quickly by pillaging their unfortunate subjects. The republic could not solve these and other problems and was in the end superseded by the monarchy of Caesar Augustus.

The empire. The bedrock of Augustus' power was his command of the legions, but he himself was a much better politician than he was a general, and he knew quite well that naked political power is as unstable as it is expensive. He reduced the military establishment as much as was prudent, laboured to turn the revolutionary faction that had supported his bid for power into a respectable new ruling class, and proclaimed the restoration of the republic. But not even Augustus could make the restoration real. With the safety of the state, questions of war and peace, and most of the business of governing the empire in the hands of a monarch, there was not enough for the Senate to do, and Augustus never went so far as to restore genuinely free elections or the organs of popular government. He kept the crowd happy with chariot races, gladiatorial contests, and the dole of bread. Nevertheless, he could not give up the attempt to legitimize his regime. Like earlier monarchs elsewhere, he called in the aid of religion, even though the religion of Rome was as republican as its constitution. Later emperors made their own divinity a tenet of the public faith.

The Augustan state

For four centuries the resemblance between Rome and the bureaucratic eastern monarchies steadily increased. Roman nationalism, Roman traditionalism, and Roman law survived as legacies that posterity would one day claim; and if nobody much believed in the constitutional shams of Augustus' day, the example of his constitutional monarchy was to prove potent at a much later period.

The age of the city-state was at last drawing to a close. The emperor Caracalla extended Roman citizenship to all subjects of the empire, but that was merely so that he could tax them more heavily. The demands of the imperial administration began to bankrupt the cities, which had previously prospered as the local organs of government under Rome. New barbarian attacks threw the empire onto the defensive, and in AD 410 Rome fell to the Visigoths.

THE MIDDLE AGES

Dissolution and instability. Seen against the background of the millennia, the fall of Rome was so commonplace an event that it is almost surprising that so much ink has been spilt in the attempt to explain it. The Visigoths were merely one among the peoples who had been dislodged from the steppe in the usual fashion. They and others, unable to crack the defenses of Sāsānian Persia or of the Roman Empire in the east (though it was a near thing), probed farther west and at length found the point of weakness they were seeking on the Alps and the Rhine.

What really needs explaining is the fact that the western empire was never restored. Elsewhere the universal throne was never vacant for long. In China, after every time of troubles, a new dynasty received "the mandate of Heaven," and a new Son of Heaven rebuilt order. For instance, in AD 304 the Huns had invaded China, and a long period of disruption followed; but at the beginning of the 7th century the T'ang dynasty took charge and began 300 years of rule. The Europeans failed to emulate this story. Justinian, the greatest of the eastern Roman emperors, reconquered large portions of the West in the 6th century, though the destruction wreaked by his soldiers made things worse rather than better. In 800, Charlemagne, king of the Franks, was actually crowned emperor of the West by the Pope. In later centuries the dynasties of Hohenstaufen and Habsburg tried to restore the empire; so, as late as the 19th century, did Napoleon Bonaparte. None of these attempts succeeded. Probably the chance was only real in the earliest period, before Europe had got used to doing without an overlord. But at that time there was never a long enough breathing space for society to regain its stability and strength. Most of the barbarian kingdoms, successor states to Rome, succumbed to later assailants. Britain fell away from the empire in the 5th century; the little kingdoms of the Angles and Saxons were just coming together as one kingdom of England when the Viking invasions began. In the 7th century the Arabs conquered North Africa, and in the 8th they took Spain and invaded Gaul. Lombards, Avars, Slavs, Bulgars, and Magyars poured into Europe from the East. Not until Otto I's victory over the Magyars at Lechfeld in 955 did these incursions cease, and not until the late 11th century was Latin Christendom more or less secure within its borders; and by then it had been without an effective emperor for more than 600 years.

Attempts at restoration

Feudalism. Various institutions had emerged to fill the gap. The church, against enormous odds, had kept the light of religion and learning alive and spread what was left of Roman civilization into Ireland, England, Central Europe, and Scandinavia. It also provided a reservoir of literacy against the day when professional government

should be possible again. The kings of the barbarians, of whom Charlemagne was the greatest, had provided military leadership and tried to acquire some of the prestige and governmental machinery of the Roman emperors. But the troublous times ensured that effective power fell to a military aristocracy. Its members called themselves *nobiles* in the Roman fashion and appropriated various late imperial titles such as *comes* (count) and *dux* (duke). But this was mere decoration. The new kings, lacking the machinery for imperial taxation, could not pay for standing armies. Besides, this was the age in which the heavily armoured cavalryman (*chevalier* in French; knight in English) dominated war. He was an independent force and thus a much less dependable instrument than a Roman legionary had been. Legally, the new masters of the soil were liegemen of the various kings (it was a maxim that every man had a lord), but in practice they could usually ignore royal claims. Europe thus fell under the rule of armoured lords; and the course of the next few hundred years gave everyone reason to wonder whether the democrats of Greece had not been right to distrust the very idea of oligarchy, for the tonic note of noble rule seemed to be almost incessant warfare.

Influence of knighthood

Even at their height the military aristocrats never had it all their own way. Strong monarchies gradually developed in England, France, and, a little later, in the Iberian Peninsula. During the heyday of the papacy (*c.* 1050–1300) the Roman Catholic Church was able to modify, if not control, baronial behaviour. Trade gradually revived and brought with it a revival of the city-state in Italy, the Rhineland, and the Low Countries, for the newly prosperous burghers could now afford to build stout walls around their towns, and it became difficult for the nobility to muster the force to besiege them successfully. Even the peasants from time to time made themselves felt in bloody uprisings. The nobility itself was far from a homogeneous or united class.

The rise of law and the nation-state. Medieval Europe, in short, was a constantly shifting kaleidoscope of political arrangements; to the extent that it ever settled down, it did so on the principle that since everybody's claim to power and property was fragile and inconsistent with everybody else's, a certain degree of mutual forbearance was necessary. This explains the great importance attached to custom or (as it was called in England) common law. Disputes were still often settled by force, especially when kings were the disputants, but the medieval European became almost as fond of litigation as he was of battle. Every great estate was hung about with quasi-permanent lawsuits, and the centralization of the church on the papal courts at Rome ensured yet more work for lawyers, the greatest of whom began to merge with the military nobility into an aristocracy of a new kind. Rights, titles, and privileges were forever being granted, revoked, and reaffirmed. Parchment deeds came to regulate men's political, social, and economic relationships at least as much as the sword did. In these ways the idea of the rule of law was reborn. By the beginning of the early modern period, legally demonstrable privileges had become the universal cement of European society. The weak were thus enabled to survive alongside the strong, and Europe, where everyone knew to which order of society he belonged, thus took on a faint resemblance to India, where the caste system was so strong as to make many of the usual functions of government quite unnecessary.

The common law

But there was a dynamism in European society that prevented it from setting permanently into this or any other pattern. This evolving Europe of privileged orders was also the Europe of rising monarchies. With many setbacks the kings clawed power to themselves; by 1500 most of them presided over bureaucracies (initially staffed by clerics) that would have impressed any Roman emperor. But monarchy could no longer claim universality. The foundations of the new monarchies were purely territorial. The kings of England, France, and Spain had enough to do to enforce their authority within the lands they had inherited or seized and to hammer their realms into some sort of uniformity. This impulse explains the wars of the English against the Welsh, Scots, and Irish; the drive of the French kings toward the Alps, the Pyrenees, and the Rhine; and the rigour of the Spanish kings in forcing Catholicism on their Jewish and Moorish subjects. Uniformity paved the way for the most characteristic governmental form of the modern world, the nation-state.

EMERGENCE OF THE MODERN WORLD

The failure of absolutism. That evolution was not easy, for kings or anybody else. The invention of gunpowder enabled kings to overbear their turbulent nobles, for cannon were exceedingly effective at demolishing the castles in which rebellious barons had formerly been quite safe. But artillery was cruelly expensive. This fact underlies the rise and persistence of the greatest political discovery of the later Middle Ages: the principle of representation. A sufficient revenue had always been one of the chief necessities of monarchy; none of the great European kingdoms had ever succeeded in securing one for long. The intractable complexities of medieval society permitted very little coercion of taxpayers; for the rest, money could be secured only by chicanery, or by selling offices or crown lands (at the price of a long-term weakening of the monarch), or by robbing the church, or by a lucky chance, such as the acquisition of the gold and silver of Mexico and Peru by the king of Spain, or by dealing, on a semi-equal footing, with parliaments (or estates, as they were more generally known).

The principle of representation

Yet the principle of parliamentary representation was of slow and uncertain growth, except perhaps in England, where the folly of the Stuart kings ensured its survival. Before then, three great occurrences, the Renaissance, the Reformation, and the discovery of America, had transformed Europe.

The impact of the Renaissance defies summary, even if its political consequences are all that need be considered. The truest symbol of its importance is the printing press. For one thing, this invention enormously increased the resources of government. Laws, for example, could be circulated far more widely and more accurately than ever before. But more important still was the fact that the printing press increased the size of the educated and literate classes. Renaissance civilization thus took a quantum jump, acquiring deeper foundations than any of its predecessors or contemporaries by calling into play the intelligence of more individuals than ever before. But the catch (from a ruler's point of view) was that this development also brought public opinion into being for the first time. Not for much longer would it be enough for kings to win the acquiescence of the nobility and the upper clergy; a new force was at work, as was acknowledged by the frantic attempts of all the monarchies to control and censor the press.

The Reformation was the eldest child of the press; it, too, had diffuse and innumerable consequences, the most important of which was the destruction of the Roman Church's claim to universality. It had always been a somewhat fraudulent claim: the pope had never actually been accepted by all the Christian bodies, even all the orthodox ones; but after Martin Luther and John Calvin, the scope of his commands was radically reduced. In the long run the consequence was the secularization of politics and administration and the introduction of some measure of religious toleration. Gradually the way became clear for rational, utilitarian considerations to shape government.

The discovery of the Americas opened a new epoch in world history. The Spanish overthrew the monarchies of the Aztecs and the Incas, thanks partly to the superior weapons that they brought with them and partly to the diseases. It was a spectacular episode, the first to proclaim that the old struggle between the steppe and the sown had been bypassed: from now on the drama of history would lie in the tension between the oceans and the land. The globe was circumnavigated for the first time; European ships, bearing explorers, traders, pirates, or men who were something of all three, penetrated every sea and harbour; and although the ancient civilizations of Islām, India, China, and Japan saw no need to alter their customs to take account of European innovations, the signal had been given for their fall.

Portuguese and Spanish explorations gave far-flung overseas empires to both countries, and perhaps as many difficulties as benefits. Other countries—France, England, the Netherlands, Sweden, Denmark—thought it both undesirable and unsafe not to seek empire themselves; the Iberian monarchies were thus involved in a perpetual struggle to defend their acquisitions. This entailed incessant expenditure, more, in the end, than the kingdoms' revenues could match. Financial weakness was one of the chief causes of the ultimate decline of Spain. But by then the inadequacies of the monarchical system had been cruelly exposed in such episodes as the revolt of the Netherlands, the defeat of the Spanish Armada by England, and, worst of all, the snail's-pace development of the colonies in the New World. Charles V and Philip II were as able as all but a few monarchs in history; but they could not overcome the structural weaknesses of hereditary monarchy. There was no mechanism by which they could devolve their most crushing duties on their ministers, so government moved slowly, if at all. As lawful sovereigns they were bound by the customs of their numerous realms, which frequently blocked necessary measures. They were unable to guarantee that their heirs would be their equals. The same difficulties eventually ruined the French monarchy. The only remedy discoverable within the system was for the king in effect to abdicate in favour of a chief minister. Unfortunately a man equal to the task was seldom found, and anyway no minister, however great, was ever safe from the constant intrigues and conspiracies of disgruntled courtiers. Problems tended to accumulate until they became unmanageable.

Weaknesses of monarchy

Representation and constitutional monarchy. Meanwhile, in England, the rise of Parliament introduced a republican, if not a democratic, element into the workings of one of Europe's oldest kingdoms.

The republican tradition had never quite died out. The Dutch had emerged from their long struggle against Catholic Spain clinging triumphantly to their new religion and their ancient constitution, a somewhat ramshackle federation known as the United Provinces. Switzerland was another medieval confederation; Venice and Genoa were rigidly oligarchical republics. What set England apart was that everywhere else in monarchical Europe the medieval representative institutions—the États-Généraux in France, the Cortes in Spain, the various diets of the Germanic countries—fell into oblivion. In England alone the expense of naval war, the incompetence of the kings, the exhaustion of all other sources of finance, and the absence of a standing army brought about the continuance and strengthening of Parliament. The climax came under William III, a Dutchman who was quite content to let Parliament take an unprecedentedly large share in government so long as it voted money for his war against Louis XIV of France. He conceded, in short, full power of the purse to the House of Commons, and before long it became a maxim of the dominant Whig Party that no man could be legally taxed without his own consent or that of his representatives. A radically new age had dawned.

The Whig system was called constitutional monarchy. Disputes about the right and legal way of governing England had raged throughout the 17th century. Civil wars and revolutions had been accompanied, perhaps in part caused, by endless rummaging through old manuscripts and a nearly ceaseless war of printed pamphlets. The increasingly rationalist temper of the times, exemplified in the works of John Locke, finally buried some of the more blatantly mythological theories of government, such as that of the divine right of kings; and after the flight of King James II in 1688, Parliament finally settled the issues that had so vexed the country by passing a series of measures that gave England a written fundamental law for the first time. Henceforth the country was to be ruled by a partnership between King and Parliament (in practice, between the king and an oligarchy of country gentlemen who controlled most parliamentary elections); and if many Englishmen looked with distaste on the squabbles of party politics, which were the sordid result of this arrangement, few could propose a plausible alternative. Tories drank toasts to the King over the water; republicans published more pamphlets; and Sir Robert Walpole ruled for 21 years as the first British prime minister.

Rise of the British Parliament

The secret of Walpole's strength lay in his ability simultaneously to please the King, to give the country sound government finances, and to command a majority in the houses of Parliament. He performed this last trick partly by giving out sinecures, salaries, and titles to his supporters, partly by his superiority in debate, and partly by exploiting Whig fears of Tories and Roman Catholics. These three elements—party interest, practical decision making, and party ideology—have in one form or another come to dominate most modern political systems where brute force is at a discount.

Even after Walpole's fall his arrangements continued. They were vindicated by the Seven Years' War (1756–63), when Britain defeated both the French and Spanish empires and emerged predominant in every ocean and (especially) in North America. Immediately afterward republican ideology found its classical expression.

The American and French republics. The limited British monarchy found it little easier to govern a sea-borne empire than did the kings of France and Spain. If Britain's American colonies were to grow in population and riches, so as to become sources of strength to the empire, not military and financial liabilities, they had to be given a substantial measure of religious, economic, and political independence, and the gift could not be revoked. Once British policy had created a chain of more or less self-governing communities along the Atlantic seaboard of North America—communities much like the city-states of old—it could not undo its own work, even when it found its clients unreasonable, small-minded, and recalcitrant. Thus, when the British government attempted to impose tighter rule from London, the old empire broke down in bickering about taxation and, in riot, rebellion, and civil war—in short, in the American Revolution. From 1775 to 1783 the Anglo-Americans fought with determination and good luck against their former overlord, King George III; and in 1776 their leaders determined to be free of him and the British Parliament forever. The principles on which they meant to found a new commonwealth were expounded in the Declaration of Independence:

The Declaration of Independence

> . . . We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain inalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. —That, to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed. . . .

The application of these principles was more difficult than their enunciation, and it took the Americans 12 years to get themselves a suitable form of government. When they did adopt a new constitution it served them so well that it is still operative. This durability is not unconnected with the fact that the Constitution opened the door to modern liberal democracy. "The consent of the governed" was agreed to be the key to governmental legitimacy, and in practice the phrase rapidly came to mean "the consent of the majority." The principle of representation was embodied in the Constitution (the first article of which was entirely devoted to the establishment of Congress, the American parliament); this implied that there was no limit to the potential size of a successful republic. From Plato to Rousseau, theorists had agreed that democracies had to be small, because by definition all their citizens had to be able to give their consent in person. This restriction was alone almost sufficient to explain the recurrent failures of republican government down the ages. Now it had been evaded.

Yet its example might have had little effect on Europe but for the French Revolution of 1789. The French had helped the Americans to beat the British, but the effort had been too much in the end for the monarchy's finances. To avert state bankruptcy the États-Généraux were summoned for the first time in 175 years, and soon the whole government had been turned upside down. The French repudiated the divine right of kings, the ascendancy of the nobility, the privileges of the Roman Catholic Church, and the regional structure of old France, and, in the end, they set up a republic and cut off the former king's head.

Unfortunately for peace, in destroying the monarchy, the Revolution also crowned its centuries-old labours. The kings had created the French state; the Revolution made it stronger than ever. The kings had united their subjects in the quest for glory; now the nation made the quest its own. In the name of rationality, liberty, and equality (fraternity was not a foremost concern), France again went to war. The Revolution had brought the new invention, the nation-state, to maturity, and soon it proved capable of conquering the Continent, for everywhere French armies went, the revolutionary creed went, too.

The nation-state

In all this the French Revolution was giving expression to a general longing for government devoted to the greatest happiness of the greatest number. But there was also considerable resistance, which increased as time went on, to receiving the benefits of modern government at the hands of the French. Besides, the facts of demography began to tilt against the Revolution, as population growth in Britain and Germany accelerated and that of France slowed down. The century that began with the victories of Marengo, Austerlitz, and Jena ended with the Third Republic nervously on the defensive and French society still bitterly at odds with itself.

Yet, on the whole, the work of the French Revolution survived. However many changes of regime France endured (seven between 1814 and 1870), its institutions had been thoroughly democratized, and the underlying drift of events steadily reinforced this achievement. By mid-century universal manhood suffrage had been introduced, putting France in this respect on the same footing as the United States. Britain, pursuing its own historical logic, evolved in much the same way; its oligarchs slowly and ungraciously consented to share political power with other classes rather than lose it altogether. By the end of the century manhood suffrage was clearly at hand in Britain, too, and women would not much longer be denied the vote. Smaller European countries took the same course. And everywhere (in America as well as western Europe) the representative principle combined with the necessity of government to produce the modern political party. Elections could be won only by organized factions; politicians could attain or retain power only by winning elections. Permanent parties resulted. The Industrial Revolution and continuing population growth made an elaborate state apparatus increasingly more necessary. The spread of education and prosperity made more citizens feel fully equal to taking part in politics, whether as voters or statesmen. Modern government in the West thus defined itself as a blend of bureaucracy, party politics, and passionate individualism, the whole held together by the cement of an equally passionate nationalism.

Political parties

Nationalism and imperialism. The kingdom of Prussia and the empires of Austria and Russia readily learned from the French Revolution that it was necessary to rationalize government. They had been struggling along that path even before 1789. Carrying out the necessary changes proved exceedingly difficult. (Russia, which in some ways was more like ancient Egypt than a modern country, made far too few changes until far too late.) Meanwhile the libertarian and egalitarian components of the revolutionary legacy were rigidly resisted. The great dynasts, and the military aristocracies that supported them, had no intention of admitting their obsolescence. Though they were forced to make limited concessions between 1789 and World War I, the autocratic citadel of their power was never surrendered. Instead, the myth of the nation was adopted to reinforce the authority of the state.

Nationalism intensified the competitiveness that had always been a part of the European state system. Peoples, it emerged, could be as touchy about their prestige as monarchs. But for a hundred years an uneasy peace prevailed in Europe, leaving the powers free to pursue interests in other parts of the world. Asia and Africa thus came to feel the full impact of European expansion, as the Americas had felt it before. Only the Japanese proved to have the resources to adapt successfully to the new ways, taking what suited them and rejecting the rest. They kept their millennial sacred monarchy but modernized the armed forces. In 1895 they fought and won a war against China, which was sliding into chaos, and in 1905 they defeated a great power, Russia. However, Japan was wholly exceptional. Elsewhere European power was irresistible. Britain gave up the attempt to govern overseas settlements of its own people directly—the experiment had proved fatal in America and nearly so in Canada—but it had no scruple about assuming direct rule over more non-British peoples. France, Germany, and the United States eagerly followed the example; The Netherlands, Spain, and Portugal clung to what they had, though the last two suffered great imperial losses as Mexico, Brazil, and others shook off colonial rule. It seemed that before long the whole world would be ruled by half a dozen powers.

It did not turn out so, or not for long. The problem of governmental legitimacy in central, eastern, and southern Europe was too explosive. The obstinate conservatism of the dynasts proved fatal to more than monarchy. There were too many who regarded the empires as unacceptable, either because they were the instruments of class oppression, or because they embodied foreign rule, or both. And the romantic tradition of the French Revolution—the fall of the Bastille, the Reign of Terror, the Jacobin dictatorship—helped to drive many of these critics into violent rebellion, permanent conspiracy, and corrosive cynicism about the claims of authority. Authority itself, corrupted by power and at the same time gnawingly aware of its own fragility, embarked on more risky militarist adventures. The upshot was the war of 1914, the revolutions that followed it, and the new model government that emerged in the Russian empire at the end of the period.

Legitimacy and upheaval

20TH-CENTURY MODELS

The Soviet state. In theory the Soviet Union was a democracy, but in practice it was an oligarchical tyranny. Lenin and his followers had won power in the turmoil of revolutionary Russia because they were abler and more unscrupulous than any other group. They retained and increased their power by force, but they argued that the social theories of Karl Marx, as developed by Lenin, were of universal, permanent, and all-sufficient validity, that the leadership of the Communist Party had a unique understanding of these theories and of the proper tactics for realizing them, and that therefore its will could never legitimately be resisted. All institutions of the Soviet state were devised primarily to assure the untrammelled power of the party, and no methods were spurned, from mass starvation to the suppression of works of art, in furtherance of this aim. Even the economic and military achievements of the regime were secondary to this overall purpose; its most characteristic institution, next to the party, was the secret police.

The Soviet model found many imitators. Lenin's strictly disciplined revolutionary party, the only morality of which was unswerving obedience to the leader, was a particularly attractive example to those intent on seizing power in a world made chaotic by the world wars. Benito Mussolini of Italy modelled his Fascist Party on the Leninists; Adolf Hitler of Germany copied Mussolini; Mao Zedong led a Communist movement in China. All three men won power, which they criminally abused, in conditions of social and political collapse like those that Lenin had exploited. In the Soviet Union itself, Lenin's successor and disciple, Joseph Stalin, outdid his master in building up his power by mass terror and party discipline. After World War II, Stalin was able to extend the Soviet model directly to eastern Europe. However, after his death in 1953 his empire began to lurch from crisis to crisis, never resolving its basic dilemma: party dictatorship condemned the Soviet Union and its vassal states to permanent inefficiency and unrest, but reform would probably blow away the Communist ascendancy. In 1985 a new generation came to power with Mikhail Gorbachev, who was willing to take enormous risks in order to revitalize the Soviet empire. Before long the Communist regimes disintegrated in most of central and eastern Europe, and in 1991 the Soviet Union itself dissolved. In China a similar attempt at reform ran into similar difficulties. It thus became clear suddenly to most of the world that the Leninist experiment had definitively failed.

Stalinist totalitarianism

Liberal democracy. Government in the West after World War II evolved more successfully. The democratic system was everywhere in the ascendant and brought with it greater influence for the working classes, for women, for non-European races, and for small states. The turbulent processes of open debate and decision produced an economic order that proved vastly more productive than the command economies of the East and brought greatly improved living standards for the great majority. Free elections meant that bad, or at any rate unpopular, governments were regularly dismissed by the voters with historically unparalleled ease and peacefulness. No Western democracy made war against any other.

But Western democracy, however perfect its forms (and nowhere were they entirely consistent with its principles), always had problems on its hands that might in the end prove too much for it. The Soviet threat might be replaced by others, as was dramatically demonstrated in 1990 when Iraq seized the emirate of Kuwait. Class conflicts were muted rather than resolved. Nationalism still distorted voters' judgments in matters of foreign policy; greed misled them over economic policy. Demagogues abounded as much as they ever did in ancient Athens, and many politicians were corrupt.

The great experiment of European imperialism had collapsed. The two world wars robbed the powers of the will and the means to maintain overseas rule. Unfortunately, the empires were not immediately succeeded by new governmental forms fully capable of dealing with the problems of technologically backward, overpopulated, culturally premodern societies. In the Muslim world the idea of the Islāmic republic arose in the 1980s. In Iran it amounted to an attempt to wed religion and government indissolubly, the religion being Shī'ite Islām. This Islāmic Leninism has not yet been tried anywhere else, and in any case it cannot spread outside regions of Islāmic dominance. In India a regime of more or less democratic nationalists endeavours to overcome a pervasive regionalism and social stratification that have been part of the subcontinent for most of its history. Japan has adapted to Western notions of capitalism and parliamentarism without contributing any fresh ideas. In the Third World the commonest expedient adopted was dictatorship, usually military, in which the ancient tradition of autocracy reasserted itself; but it was autocracy without its ancient stability, and of all current political forms it seems by far the least likely to deal effectively with the universal enemies: hunger, war, poverty, disease, waste, and violence. Some thinkers believe that only a form of world government can make decisive headway against these evils, but no one has yet suggested either how world government can be set up without a world war or how, if such a government did somehow come peacefully into existence, it could be organized so as to be worthy of its name. Even effective global cooperation among national governments is extremely difficult, as the story of such international bodies as already exist demonstrates all too sadly. In the circumstances it is wonderful that the one really important post-1945 innovation, the confederation of Europe, has made as much progress as it has.

The Islāmic republic

Another hopeful development was the ending of the ruthless competition between the two superpowers, the former Soviet Union and the United States; but the cessation of the Cold War provoked a revival of nationalism, especially in eastern Europe. Self-determination has become the universal aspiration, and the nation-state is the universal norm; these developments, however, have not yet made for general peace. The incompatible claims of the city-states ruined ancient Greece; the modern world may yet be ruined by the claims of the nations. If man, the political animal, is to save himself and his civilizations, he cannot yet rest from seeking new forms of government to meet the ever-new needs of his times.

BIBLIOGRAPHY. Classical texts on governmental forms are widely available in numerous editions. They include PLATO, *The Republic*; ARISTOTLE, *Politics*; NICCOLÒ MACHIAVELLI, *The Prince* and *The Discourses*; THOMAS HOBBES, *Leviathan*; JOHN LOCKE, *Second Treatise of Government*; MONTESQUIEU, *The Spirit of Laws*; JEAN-JACQUES ROUSSEAU, *The Social Contract*; THOMAS JEFFERSON, *Declaration of Independence*; ALEXANDER HAMILTON, JAMES MADISON, and JOHN JAY, *The Federalist*; EDMUND BURKE, *Reflections on the Revolution in France*; THOMAS PAINE, *Rights of Man*; ALEXIS DE TOCQUEVILLE, *Democracy in America*; WALTER BAGEHOT, *The English Constitution*; JAMES BRYCE, *The American Commonwealth*; FRIEDRICH ENGELS, *The Origin of the Family, Private Property, and the State*; and V.I. LENIN, *The State and Revolution*.

For a subject of this nature, a solid background in world history is absolutely necessary, and WILLIAM H. MCNEILL, *A World History*, 3rd ed. (1979), is an excellent introduction. GORDON CHILDE, *What Happened in History*, rev. ed. (1954, reprinted 1982), is a classic survey of the contribution of archaeology to our understanding of prehistory and the ancient world. For the development of governmental forms in Greece, see *The Cambridge Ancient History*, 3rd ed. (1970–); and N.G.L. HAMMOND, *A History of Greece to 332 B.C.*, 2nd ed. (1967, reprinted 1981). A.H.M. JONES, *Athenian Democracy* (1957, reprinted 1975), is indispensable, particularly as a corrective to Plato, Aristotle, and Thucydides. The most exciting work about the Romans written since Gibbon is RONALD SYME, *The Roman Revolution* (1939, reprinted 1974); those wanting more general accounts of the Romans may turn to J.P.V.D. BALSDON (ed.), *The Romans* (1966); and H.H. SCULLARD, *From the Gracchi to Nero*, 5th ed. (1982). A.H.M. JONES, *The Later Roman Empire*, 2 vol. (1964), is the most authoritative account of the fall of the empire.

The best introduction to the medieval papacy is WALTER ULLMANN, *A Short History of the Papacy in the Middle Ages* (1972). The same author's *Growth of Papal Government in the Middle Ages*, 3rd ed. (1970), links ideas and institutions but makes few concessions to beginners. MARC BLOCH, *Feudal Society* (1961, reissued 1974; originally published in French, 1939), is an indispensable study of its subject.

The development of political thought from the Renaissance to the 19th century is well presented in JOHN P. PLAMENATZ, *Man and Society*, 2 vol. (1963, reprinted 1972–74). An attempt to trace modern government back to its origins is demonstrated in BARRINGTON MOORE, JR., *Social Origins of Dictatorship and Democracy* (1966). See also PERRY ANDERSON, *Lineages of the Absolutist State* (1974); and FRANCO VENTURI, *Utopia and Reform in the Enlightenment* (1970; originally published in Italian, 1970).

Works on governmental forms of a general, comparative nature include A. GOODWIN (ed.), *The European Nobility in the Eighteenth Century*, 2nd ed. (1967); A. LAWRENCE LOWELL, *Governments and Parties in Continental Europe*, 2 vol. (1896, reissued 1970); EUGENE N. ANDERSON and PAULINE R. ANDERSON, *Political Institutions and Social Change in Continental Europe in the Nineteenth Century* (1967); MICHAEL OAKESHOTT (ed.), *The Social and Political Doctrines of Contemporary Europe* (1939, reissued 1949); and S.E. FINER, *Comparative Government* (1970, reissued 1974).

Useful studies of particular states include J.H. PLUMB, *The Growth of Political Stability in England, 1675–1725* (1967, reissued 1980: also published as *The Origins of Political Stability, England, 1675–1725*); and ÉLIE HALÉVY, *A History of the English People in the Nineteenth Century*, 6 vol. (1924–34; originally published in French, 5 vol., 1913–32); for France, C.B.A. BEHRENS, *The Ancien Régime* (1967, reissued 1976); GEORGE LEFEBVRE, *The French Revolution*, 2 vol. (1962–64; originally published in French, 1930); and D.W. BROGAN, *The French Nation from Napoleon to Pétain* (1957, reissued 1961); for Germany, HANS ROSENBERG, *Bureaucracy, Aristocracy, and Autocracy: The Prussian Experience, 1660–1815* (1958, reprinted 1968); A.J. NICHOLLS, *Weimar and the Rise of Hitler*, 2nd ed. (1979); FRANZ NEUMANN, *Behemoth*, 2nd ed. (1944, reissued 1963); and MARTIN BROSZAT, *The Hitler State: The Foundation and Development of the Internal Structure of the Third Reich* (1981; originally published in German, 1969); for Italy, DENIS MACK SMITH, *Italy: A Modern History*, new rev. ed. (1969); and S.J. WOOLF (ed.), *The Nature of Fascism* (1968); for the Soviet Union, JEROME BLUM, *Lord and Peasant in Russia: From the Ninth to the Nineteenth Century* (1961, reprinted 1971); LEONARD SCHAPIRO, *The Origin of the Communist Autocracy: Political Opposition in the Soviet State, First Phase 1917–1922*, 2nd ed. (1977); and DEREK J.R. SCOTT, *Russian Political Institutions*, 4th ed. (1969); and, for the United States, RICHARD HOFSTADTER, *The American Political Tradition and the Men Who Made It* (1948, reprinted 1974), and *The Idea of a Party System: The Rise of Legitimate Opposition in the United States, 1780–1840* (1969). (H.Br.)

Government Finance

Government finance refers to the wide range of activities undertaken by governments in financial and economic matters. In earlier times, a government's financial activities were limited to raising funds, whether from taxes or other levies, to support its wars and its other activities. For centuries sovereigns were allowed to spend such funds as they chose; the control of expenditures came much later, not becoming a matter of concern for legislatures until the late 17th and 18th centuries. The tradition grew up that expenditures must flow from appropriations, and appropriations were to be made for specific purposes. Out of this evolved the more formal governmental budgets of today.

As the proportion of national income that is controlled by the various levels of government has increased, so has the importance of the budget for the general working of the economy. If a government raises more money than it spends (creating a surplus), the overall level of spending in the economy—along with personal income and employment—may be reduced. Conversely, if the government spends more than it raises, which creates a deficit, fiscal and monetary expansion is implied, which may influence inflation and growth. Thus budgets have come to serve as instruments of economic management, used to stabilize or stimulate the economy. The composition of the budget also has other economic effects. The most important of these effects are brought about by its allocative function (deciding what is supplied by the public sector and to whom) and its distributive function (deciding who gets what and how income is distributed and redistributed). The national budget, then, has become the most important instrument of a nation's economic management, and in many countries the presentation of the new budget has become synonymous with the primary annual assessment and statement of economic policy. For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 531–536.

This article is divided into the following sections:

The governmental budget 148
- Role of the budget 148
  - Traditional functions
  - Modern functions
- The accounting functions of the budget 149
- Alternative approaches to the budget 149
  - Administrative budget
  - Current and capital budget
  - Cash and unified budgets
  - Program budgeting
  - Full-employment budget
  - Value for money measurements
  - Budgetary planning: cash, volume, and cost terms

Components of the budget 151
- Expenditure 151
  - Composition of public expenditure
  - Growth of public expenditure
  - Problems of public expenditure control
- Revenue 152
  - Taxation
  - Sale of goods and services
- Government borrowing 153
  - Forms of public debt
  - Economics of government borrowing
  - Limitations on public sector debt
  - Wartime finance
  - Evolution of government borrowing
  - Sovereign debt

The economic functions of government 157
- The allocative function 157
  - Public goods
  - Merit goods
  - Cost-benefit analysis
  - Public ownership and privatization
  - Other forms of government intervention
- The stabilization function 158
  - History of stabilization policy
  - Stabilization theory
  - Fiscal policy
  - Monetary policy
  - Stabilization policy problems
- The distributive function 159

Bibliography 159

The governmental budget

A governmental budget is the forecast by a government of its expenditures and revenues for a specific period of time. In national finance, the period covered by a budget is usually a year, known as a financial or fiscal year, which may or may not correspond with the calendar year. The word budget is derived from the Old French *bougette* ("little bag").

ROLE OF THE BUDGET

Traditional functions. Government budgetary institutions in the West grew up largely as a result of the struggle for power between the legislative and executive branches of government. With the decline of the feudal system, it became necessary for kings and princes to obtain resources for their ventures from taxation rather than dues. With the disappearance of the old feudal bonds, taxpayers demanded to be consulted before they were taxed. In England this was written into Magna Carta (1216), which stated:

> No scutage or aid shall be imposed in our kingdom unless by common counsel of our kingdom, except for ransoming our person, for making our eldest son a knight, and for once marrying our eldest daughter, and for these only a reasonable aid shall be levied.

This related to taxes only, not expenditures. For centuries Parliament seemed content to restrict the amounts that the sovereign levied while letting him spend the money as he pleased. Only after the controversies of the 17th century culminated in the Revolution of 1688 and the Bill of Rights did Parliament extend its concern from taxation to the question of expenditure control.

The histories of many countries have turned on financial crises. In France, for instance, the struggle between the monarchy and the nobility over control of tax revenues was one of the causes of the Revolution of 1789 that led to the overthrow of both the monarchy and the nobility.

The U.S. budget system also evolved out of controversy. In the early days of the republic there was a dispute between Alexander Hamilton and Thomas Jefferson as to the amount of discretion that the executive branch should exercise in the spending of public funds. Jefferson's victory enabled Congress to assert its authority by making appropriations so highly specific as to hinder executive action. Had Hamilton won, the treasury would have attained extraordinary power in relation both to Congress and to the president.

Modern functions. A high proportion of economic activity is controlled, directly or indirectly, by various levels of government (federal, or central, state, local, etc.). Thus the budget has taken on a number of other functions as well as the simple monitoring of the overall revenue and expenditure of government. Expenditure programs are now planned in considerable detail, but the sheer scale of

public spending raises major control problems, and varying systems of control have been tried in different countries. Taxation is used not only to raise revenue but also to redistribute income and to encourage or discourage certain activities. Government borrowing, in order to finance recurring deficits or wars, is so substantial that budgetary policy has important effects on capital markets and on interest and credit generally.

The budget has also come to be used to achieve specific goals of economic policy. It was long recognized that government borrowing and the levels of expenditure and taxation could affect the total demand for goods and services in the economy. This raised the possibility that by changing these levels the government could use its fiscal policy to achieve full employment and reduce economic fluctuations. This stabilization function has been used by many countries, with varying degrees of success, to expand the economy out of recession and to control inflationary pressures.

The stabilization function

As well as affecting the overall economy, the budget may have significant (intended and unintended) effects in specific areas. Taxes affect incentives to work or to consume, while taxes, benefits, and expenditures all affect the distribution of income. In this manner, budgets, particularly those that cause major changes, have considerable political as well as economic impact.

THE ACCOUNTING FUNCTIONS OF THE BUDGET

Traditionally the budget is presented to allow scrutiny (by taxpayers, voters, and the legislature) of the resources raised by government and the uses to which these will be put. The publication of a budget thus performs the role of generating accountability for the actions of government at various levels. Historically, the focus of budgets has been to ensure that expenditures and revenues are properly authorized; more recently, the budget has been developed as a framework within which complex decisions on the allocation of resources can be made more effectively.

ALTERNATIVE APPROACHES TO THE BUDGET

In order to deal with the increasing complexity of government's role, most countries have experimented with a variety of forms for the budget and its presentation. Among the more important of these are the administrative budget, the current and capital budget, program and zero-base budgeting, and the full-employment budget. The variety of budgeting methods is extended to the types of efficiency measures used to increase value for money and to the alternative methods of projecting expenditures in cash, volume, and cost terms.

Administrative budget. The traditional administrative budget contains the executive's recommendations concerning the raising of what Magna Carta referred to as "scutage or aid" and the disposal of it for purposes of government. This kind of budget is designed to control expenditure; accordingly, it emphasizes the salaries and tasks of civil servants rather than the results that they are supposed to achieve. The control objective of the administrative budget naturally gives rise to the doctrine that the budget should be balanced. Deficits imply irresponsibility. Surpluses imply the imposition of unwarranted tax burdens on the public.

The limitation of the administrative budget is that some important items receive less than adequate attention or are excluded from it entirely. Military procurement is one example. Neither budget offices nor appropriations committees are well equipped to scrutinize the actual procurement of ships or aircraft. Consequently, in most countries large expenditures on military items are often treated perfunctorily while the activities of civil servants receive inordinate amounts of attention. The basic weakness of the administrative budget is that it is principally concerned with whether expenditure has been properly authorized, rather than whether money has been well spent.

Weakness of the administrative budget

Moreover, the administrative budget often excludes trust funds used to finance contributory old-age and unemployment insurance; taxes are paid directly into the funds and disbursements made out of them. The theory is that the government acts as trustee for the public and that the public is protected by having its social security taxes put in a separate fund. Many countries have adopted this idea of "social insurance"; it formed the heart of Bismarck's social policy for Germany in the 1870s and of the British welfare state, founded in 1948. In most cases, however, the attempt to generate a distinct fund has failed, and "contributions" have become just another tax with expenditures on, for example, retirement pensions paid irrespective of the resources available to the fund.

Other items may be included in the budget on a net rather than a gross basis. For instance, the total receipts and expenditures of the post office or other commercial activities of the public sector usually do not appear; only the deficit or surplus does. This is justified by the theory that, first, business management is not well performed by legislative committees and, second, that so long as a business undertaking pays its way, its conduct is not a matter of public concern. The problem is that the distinction between commercial and noncommercial activities is often arbitrarily made.

Current and capital budget. The administrative budget traditionally deals only with current expenditures; in many countries, some items are regarded as inappropriate for inclusion because they finance capital expenditures or are loans to other public bodies. Such items are then included "below the line," and the traditional concept of budget balance is not applied to them; instead, it is regarded as permissible to finance them by borrowing.

"Below the line" items

Direct public works or investment in nationalized industries are regarded by most countries as suitable for loan financing on the ground that they are productive assets that will yield a revenue sufficient to cover their cost. They may do so either directly, as in the case of a toll highway, or indirectly by increasing the general economic welfare, as in the case of a free highway. If, however, there is no market in which the output of a public activity is sold, there can be no objective test of its value. Hence, governments are often tempted to classify expenditure on such assets as capital items that yield a social but no economic return (*e.g.*, free playgrounds) or a lower economic return than any private sector institution would accept (as in government support for declining industries). For this reason, distinctions between current and capital expenditures in public accounts are often viewed with suspicion.

This suspicion may be increased where, as is often the case, the rules for what is regarded as current or capital are rather indistinct. Moreover, governments have been reluctant to adopt the systematic distinction between current and capital items, or between cash flows and profit and loss accounts, or to construct a balance sheet, even though these mechanisms of monitoring receipts and expenditures are universal in private sector accounting. The federal government of the United States, for example, has resisted the idea of a capital budget, even though there was strong pressure for one in the 1930s when economists and politicians wanted to legitimize the government deficit. Among U.S. state and municipal governments, however, loan financing of public works is the regular practice for two reasons. First, those bodies are usually unable to finance their projects by current taxation; second, they do not want to finance them because the projects are generally of a long-term nature.

U.S. resistance to a capital budget

Most national governments have become accustomed to thinking in terms of national economic policies in which the amount of borrowing to be undertaken depends on current requirements for stability and growth. This makes capital budgeting less attractive, particularly if the government wishes to use the budget to supplement the national flow of savings. The more need there is to increase saving, the smaller should be the amount of government borrowing. On the other hand, government borrowing is justified when private savings tend to exceed private capital requirements.

This lack of explicit monitoring for the capital position of governments can have serious consequences when the government unwittingly takes on large liabilities or uses capital assets to finance current expenditures. Examples are provided by the growing problems in some countries in financing generous state pension schemes and the wasting of assets such as oil reserves.

Cash and unified budgets. Faced with the increasing complexity of government activities, many countries have fallen back on the idea of the cash budget, which resembles the cash flow account of a modern business. Trust fund expenditures and receipts are included, as well as cash payments and receipts involved in loan transactions. Government business undertakings such as the post office, however, are still included on a net basis.

The cash budget suffers from the defect that it is not directly tied to government decision making. Liabilities incurred do not synchronize completely with payments. This is because government expenditures result from appropriations and other forms of commitments; cash expenditures may follow appropriations and other commitments of money only after a considerable lag, notably in the case of construction and procurement. Appropriations relate to actions in the future. Expenditures result from past decisions. Both kinds of information are needed for a complete appraisal.

Time lags in cash budgets

In 1968 the U.S. government, in an effort to reduce public confusion over the large variety of budgetary concepts, adopted a so-called unified budget concept that is more logical than the cash budget but differs from it only in some details that do not materially affect the budget aggregates. The unified budget differs from the traditional administrative budget in two main ways: it includes the receipts and outlays of most funds, and it eliminates interagency transfers.

Program budgeting. Traditionally, government expenditures have been considered as inputs rather than outputs. This is because, in the classical 19th-century conception, the well-run government does not produce a marketable output. The program budget derives from this concept; it attempts, however, to classify expenditures in terms of the outputs to which they are devoted. For example, a traditional school budget would categorize expenditures in terms of teachers, books, and buildings; what came out of the process would be left to the reader's intuition or experience. The program budget, in contrast, attempts to assign expenditures to specific outputs, categorizing them according to numbers of children completing various programs.

In government, budgets have traditionally been constructed according to departments and agencies of government. This may be justified on historical or administrative grounds, but it does not necessarily correspond to the structure of activity. Every country organizes the civilian and military components of its foreign policy in separate departments, but this is frequently a serious obstacle to effective policy-making. Again, the requirements of good administration suggest that there should be a single department of agriculture. But that department's activities impinge on those of others, in both domestic and foreign policy. A budget constructed according to actual programs would cut across departmental boundaries.

Basic assumption of program budgeting

Program budgeting is an attempt to apply the economics of choice to public decision making. Its basic assumption is that explicit choice among alternative courses of action leads to better results than do other methods of decision making. At the highest governmental levels difficult choices must be made that involve the use of a portion of the nation's resources. But the same principles apply to decision making at lower levels. The problem of allocating resources within a specific field, such as health or education, is conceptually similar to that faced in drawing up the national budget.

Program budgeting also takes account of the time dimension in many government programs. New undertakings often take time to come into operation. A typical new program may have to pass through a research and development phase and an investment or construction phase before it reaches the operating phase. Alternative programs may differ considerably in this respect. The process of choosing among alternatives frequently involves trading the present against the future. One alternative may require 10 years before it yields results; another may yield smaller results but more quickly. The kinds of choices made in government often involve alternatives that cannot be measured in terms of market value. For this reason governmental decisions involve much more uncertainty than do most business decisions.

A governmental program must therefore be frequently revised in the light of unfolding circumstances. Indeed, every year should be thought of as the first year of a new program. Pervasive uncertainty also requires a high degree of flexibility and a capacity for program revision. A number of options should be held open, particularly in the development phase. Even though this may appear costly, it is less costly than commitment to a design that proves to be inappropriate because of circumstances that could not be foreseen in the early stages.

In most countries the usual procedure for deciding on government expenditure in a forthcoming year has been to assume that existing expenditure was appropriate and then to decide on incremental expenditure for each program. Such an approach means, however, that the change is likely to increase, rather than decrease, expenditure and that little attention is paid to what the full existing program actually accomplishes.

Full-employment budget. Although the idea of budget balance in the administrative budget has been the dominant consideration in the budgetary policy of most countries, it has gradually been realized that such a concept may be inappropriate when external shocks such as exchange rate movements or a world recession occur. Because varying levels of unemployment are a major reason why expenditures may change without comparable change in the public sector output, the concept of a full-employment budget has emerged. This type of budgeting is based on receipts and expenditures that would prevail under conditions of full employment. The approach views the actual expenditures and receipts for the coming year as of secondary importance; it assigns primary importance to the influence of the budget on the national economy. In time of recession a budget deficit may thus be presented as a necessary step toward achieving a balanced budget at full employment. Ideally, the budget should include estimates of expenditures and revenues at full employment, and also estimates of the same items at the anticipated level of employment. These ideas have been extensively used in the United States. An analogous procedure could be used with respect to inflation, but this idea is still far from acceptance, because governments are no less reluctant to anticipate inflation than they are to budget for unemployment.

Value for money measurements. As the emphasis in budgetary policy has shifted away from mere authorization of government spending and toward more public scrutiny of what government accomplishes, the idea of appraising value received for money spent in government finance has grown in importance. This has led to an increasing variety of measurements of public sector efficiency. In general terms, taxpayers need to be satisfied that their money is being used wisely. Because of the wide variety of items within even a single program, however, it is often difficult to identify precisely what is spent on the provision of each service, and the services that are provided rarely have well-developed private sector counterparts to act as a basis for comparison.

In some programs, governments have developed efficiency measures that relate observable facts, such as the quality of national health or the number of operations performed, to the cost of providing the service. The use of such measures is by no means widespread, however, and their basis is often open to question. The principal difficulty is that there is either no meaningful measure of the output of a public service—defense, for example—or output is complicated and multidimensional—as with education or health. The result is that any method used to measure efficiency is open to debate and challenge.

Obstacles to allocating efficiency

Attempts to control public expenditure, particularly since the mid-1970s, led to some examination of which programs should remain in the public sector.

Budgetary planning: cash, volume, and cost terms. There are three principal bases for public expenditure planning: cash, volume, and cost. The cash basis is concerned simply with the projected money expenditure on the services involved. Making such projections is difficult because what the cash expenditure will buy depends on what hap-

pens to prices over the planning period. Moreover, many public expenditures cannot be planned in cash terms, because legislation prescribes the output. Most social benefits, for example, must be paid to anyone who is entitled to receive them, and this means that the government cannot control directly the amount of the expenditure.

The volume basis is concerned with the planned output of public services. The difficulties of measuring output, however, have already been noted. More often the planning process, assuming that changes in inputs are associated with changes in outputs, operates with reference to the cost basis of programs.

All countries have an annual program of public expenditure allocation, in which those responsible for individual programs argue for greater allocations for their activities and those responsible for raising the money attempt to control the amount allocated. In practice, the results of this process depend as much on the political weight of individuals in charge of a spending program as on an objective assessment of its desirability. The normal practice is to take as a base what each program spent the previous year and then argue about incremental changes, rather than to consider each program in its totality, as is the case in an approach known as zero-based budgeting. This creates perverse incentives, however, in that departmental heads who have saved money in one area in a particular year have an incentive to spend more in other areas in order to protect next year's total budget.

The basis for most expenditure planning is therefore the number of public employees already in place and the volume of goods and services purchased in the base year. This, multiplied by base year prices, gives the input volume in the base year. In the late 20th century many countries (particularly the United Kingdom) began abandoning this approach, largely because it gave inadequate control of total expenditure. One reason for a given volume's costing too much to supply is the so-called relative price effect. This arises because goods and services bought by the public sector (labour, medical care, or defense equipment) may rise in price more quickly than commodities generally. Once this has been determined, volume can be expressed in cost terms. The relative price effect is somewhat subjective, however, because of the difficulty of measuring the quality of goods and services. Particularly in the case of health care and defense, the relative price effect will often contain the increased price of services and improved equipment, which are actually a volume increase.

The relative price effect

Cost measures, however, merely reflect the cost of a given input; controlling public expenditure in cost terms without taking full account of the relative price effect's change may lead to inappropriate volume responses or, more commonly, spiraling costs as existing input volume is maintained. Hence many countries have moved one stage further, attempting to monitor and control public expenditure in purely cash terms. The United Kingdom's public expenditure programs, for instance, are now "cash limited."

Although planning in cash has a superficial simplicity, in times of significant inflation it is not a very appropriate tool, and differential price rises may lead to a balance of expenditure provision somewhat different from the intended plan. In practice, although cash planning is presented as the base on which decisions are taken, those countries that have adopted this approach in fact allow informal flexibility in cash budgets, with volume measures being implicitly, if not explicitly, adopted.

Components of the budget

In the United States the budget for each fiscal year contains detailed information on the outlays intended by the federal government and the receipts expected, including those from trust funds. The budget also divides authorized expenditure into that which can be carried out without action by Congress and that which requires further authorization. In any year, about half of federal expenditure requires authorization from Congress; by witholding this authorization, Congress is able to force changes in the government's budgetary policy. The budget also summarizes the outstanding debt of the federal government and estimates the size of the surplus or deficit expected on the basis of the revenue and expenditure projected in the budget.

The U.S. budget is presented as a coherent whole for lengthy consideration by Congress, during which time it is often substantially revised. This joint consideration of revenue and expenditure is also common in most European countries. Practice in the United Kingdom, and in other countries with a British parliamentary tradition, continues to reflect the historical separation of revenue and expenditure. The U.K. budget consists of a number of different documents, with only limited attempts being made to relate one to another. A pre-budget report of the government's intentions is given in an Autumn Statement, usually published in November, and detailed expenditure plans are provided in February or March in a White Paper. The U.K. budget, usually presented in March, is mainly concerned with taxation and is represented in a separate volume entitled *Financial Statement and Budget Report.* This gives a general outline of budgetary strategy, details of proposed tax changes, and estimates of likely revenues, as well as details of such items as capital receipts from asset sales and the size of the contingency reserve of unallocated money to cover unforeseen events.

Variations in U.S. and U.K. budgets

Partly because of this fragmentation of the U.K. budget, and the difficulty of relating the public expenditure White Paper to the *Financial Statement and Budget Report,* debate is limited, and it is rare for any detail to be changed after the documents are published. The fragmentation of the budget is exacerbated further by the presentation of details of social security expenditure in yet another document.

EXPENDITURE

Composition of public expenditure. Expenditures authorized under a national budget are divided into two main categories. The first is the government purchase of goods and services in order to provide services such as education, health care, or defense. The second is the payment of social security and other transfers to individuals and the payment of subsidies to industrial and commercial companies. Both types are usually labeled "public expenditure," and in many countries attention usually focuses on the aggregate of the two. This obscures important differences in the economic significance of the two items, however. The first represents the public sector's claim on total national resources; the second the scale of its redistribution within the private sector.

The two main categories of public expenditure

In most Western countries, the share of the public sector in total economic activity averages between 20 and 30 percent. This reflects the proportion of workers who are employed in the public sector or in publicly financed activities, the proportion of national output generated there, and the proportion of incomes derived for productive services that is earned by public sector employees.

Some of these activities yield commercial revenues—the postal service, for example. Most have to be financed by taxation. In addition, the government raises taxation in order to redistribute income within the private sector of the economy. It taxes some activities and subsidizes others—through investment credits, for example. On a larger scale, it uses the benefit and social security system to make payments to needy individuals and raises taxes in order to subsidize those who warrant it. With this redistributive activity, plus the direct government productive activity financed from legislation, the total share of incomes taken in taxation is higher than the share of government in total production. It averages around 40 percent in Western economies.

In addition to direct expenditures, attention has been drawn to "tax expenditures." If the government favours a particular activity—such as investment—grants or tax concessions may be awarded to that activity. The two procedures have much the same effect on investment and on government revenues, but one appears to raise public expenditure and the other to reduce taxation. It has been suggested that these tax expenditures—tax reductions motivated by an economic or social objective—should be the subject of a tax expenditure budget similar to the pub-

lic expenditure budget, and several countries have now moved in that direction.

For all private and public purposes within the economy, the scale of public activity is best measured as a proportion of national income: the total of incomes generated or (equivalently) of expenditures on goods and services.

The overall proportion of national income that is collected in taxes, raised from profits on government activities, or borrowed varies widely in the developed nations. This variation reflects different national decisions concerning the proportion of a nation's activity deemed most appropriate to have carried out by the various levels of government or by government agencies. Much of the variation occurs because of choices over the provision of health care and over the level and importance of transfer payments.

Transfer expenditures

Expenditures on transfers also vary widely, depending partly on how redistributive the government wishes to be, partly on how much of this redistribution is carried out through the tax system, and partly on factors such as the number of old people and the level of unemployment. The dominant payment in every country is for old-age pensions, and the amount depends on how well-developed private sector pensions are. Another factor is the extent to which the government chooses to use direct subsidies rather than tax concessions to stimulate the economy.

(J.A.Ka./Ed.)

Growth of public expenditure. The proportion of national income devoted to public spending rose considerably during the 19th and 20th centuries. Much of this historical rise, however, cannot be taken as a direct measure of either the relative importance of government as a whole in economic decision making or of the comparative roles of central and lower levels of government. Inflation aside, in most countries the major reasons for the persistent rise in public spending since the middle of the 19th century have been war and the preparation for war, the rise in the cost of pensions for veterans, the great increase of the administrative role of government in response to expanded and urbanized populations, and the marked rise in the demand for a varied list of public services as the vote was gradually extended to the lower income classes.

Charles Bastable's observations

Writing in 1890, the Irish economist Charles Bastable observed that "in nearly all modern States outlay is steadily increasing," and "the older doctrines of economy and frugality have disappeared." He was referring to doctrines that had developed in the latter part of the 18th century, particularly in connection with the Industrial Revolution. He did not mean that there had been a "golden age" in which governments entirely refrained from interfering in the private sector. As Bastable himself pointed out, even the strictures of Anne-Robert-Jacques Turgot and Adam Smith on "excessive" government intervention did not preclude the encouragement of new industries.

In western Europe there was a long tradition of government influence on private economic decisions. The interventionist policies in the England of Henry VIII, Elizabeth I, and Oliver Cromwell, the France of Louis XIV and Colbert, and the Russia of Peter the Great are examples of such influence. But the sense of confidence conferred on the industrial class through the industrial and transportation revolutions of the 19th century, especially in Britain and the United States, produced an atmosphere that was unfavourable to government intervention. This did not, however, prevent rising pressure for government spending on economic resources, together with a secular rise in the magnitude and variety of the output of public goods that is still in evidence.

(J.A.Ka./K.E.P.)

Problems of public expenditure control. The problems of controlling public expenditure vary across programs. Some are "demand led." Transfer payments, and particularly social security payments, are largely dependent on the number of old or unemployed people. Apart from reducing benefits (which may in turn be prevented by past commitments), or through macroeconomic policies designed to reduce unemployment, for example, there is little that can be done to limit these payments. Most countries have seen a steady rise in transfer payments as the longevity of the population and the benefits of pension schemes increase.

Public expenditure also depends on the price of the goods and services that the public sector buys and on the efficiency with which they are used. Public sector workers are often highly organized and may be well placed to demand pay increases from an employer who is able to recoup the costs from taxation. Public sector purchasing may be inefficient—civil servants may find it easier to enjoy a comfortable relationship with their suppliers, and, in fields such as health and military expenditures, administrators may demand the latest technologies with little regard for their cost-effectiveness.

At the same time, much of the public sector lacks the incentives to increase efficiency that apply to private firms in competitive markets. It is easier to resist innovation, and bureaucracies often have a conservative culture in which it is more important to avoid mistakes than to experiment with new techniques and procedures. With few external indicators of performance, managers in the public sector may feel inclined simply to promote the growth of their organization and the staff numbers and budgets that they control.

As the level and complexity of governmental involvement in the economy has risen, so public expenditure has become increasingly difficult to control. The only people with enough information to monitor their program needs are those actually engaged on the program. Coupled with technological change, the general tendency has been for expenditures to rise without any clear evidence of increased levels of service being provided. Indeed, in many key areas, such as health and education, expenditures have risen steadily at the same time that the public perceived a deterioration of service.

Resistance to rising expenditures

Governments in most countries have responded to this problem by occasional severe contraction of particular programs or of public expenditure in general. Numerous countries have adopted cost-cutting exercises with some limited success. But attempts at cost reduction can provoke inappropriate reactions. If politicians discover expenditure can be reduced without reducing the value of the services provided, they may insist on further cuts. If, on the other hand, popular or politically sensitive activities are restricted, there will be pressure to restore expenditures. Managers of public sector programs therefore often have incentives to respond to cuts in ways that maximize, rather than minimize, the effects on the services provided.

REVENUE

Governments acquire the resources to finance their expenditures through a number of different methods. In many cases, the most important of these by far is taxation. Governments, however, also have recourse to raising funds through the sale of their goods and services, and, because government budgets seldom balance, through borrowing. The subject of borrowing, because of the intricacies of deficit spending, is covered in a separate section of this article.

Taxation. Most countries raise resources through a variety of taxes, including direct taxes on wage and property income, contributions to trust funds, and a variety of indirect taxes on goods, either at the final point of sale or on the inputs used to make them. A smaller amount of revenue is raised from taxes on property, on capital gains, and on capital transfers, particularly at death. Most countries have a separate corporate income tax.

The composition of tax revenues. The balance between these different taxes has varied considerably over time and between countries. In the United States, sales taxes are relatively unimportant, accruing mainly to state and local governments. Federal government revenue is principally derived from taxes on personal and corporate income. This dominant reliance on income taxes in the United States is a post-World War II phenomenon; at the beginning of the 20th century about half of all tax revenue came from taxes on property and half from sales taxes. Income tax was introduced on a regular basis only in 1913.

The tradition in Europe is somewhat different, with indirect taxes being relatively more important. All the countries in the European Communities impose a tax (at varying rates) on value added, charging tax on output from

Value-added tax

industry and rebating it on inputs. In the United Kingdom, for example, value-added tax (VAT) raises about half as much as the personal income tax, and together excise duties and VAT raise about one-third of total tax revenue.

Australia, New Zealand, and the Scandinavian countries all rely heavily on income and profits taxes, which account for about half of all revenue raised from taxation.

Payroll tax

Payroll taxes are relatively unimportant, raising significant amounts only in Australia, Austria, France, Ireland, and Sweden but rarely exceeding 5 percent of total revenue. Property taxes rarely account for more than another 5 percent, with the United Kingdom being the exception in this case. Sales taxes, excise duties, and VAT account for nearly one-half of all revenue in Greece, Ireland, and Portugal, compared with less than one-fifth in Japan.

The relationship between tax rates and revenues. In deciding how to raise enough money to finance its expenditure program, a government faces a large number of different considerations. First, the tax system is complex, containing many different taxes, each often having a complex structure. Perhaps the major consideration is the effects on behaviour that particular tax rates will cause.

Income tax has a graduated structure whereby no tax is paid on the first segment of income and then each subsequent segment is taxed at a higher rate than the previous one. In the United Kingdom most taxpayers pay tax at a uniform marginal rate, while other countries have more steeply rising rate schedules. Higher marginal tax rates make work less rewarding, which tends to reduce work effort. High marginal rates, however, may have less impact in some areas than others, a factor that needs to be considered when deciding who should bear the tax burden. Such considerations presumably have influenced the trend in many countries to tax the wealthiest groups.

Whatever the structure of the tax, the general proposition that increasing tax rates will reduce work effort usually holds; and this, in turn, tends to reduce tax revenue again. A debate arose over the "Laffer curve," which postulates that at some level of tax the disincentive effects will be so great as to mean that an increase in tax rates actually reduces revenue.

Effect of tax rates on consumption

Tax rates affect the pattern and level of consumption. Excise duties, value-added tax, and sales taxes all change the relative prices of goods and the attractiveness of consumption relative to saving. Once again, an increase in tax rates will generate responses that tend to cause a reduction in revenue, and, again, governments must balance the strength of these effects when deciding on which rates to increase. Other considerations, such as the protection of domestic industries, also affect such decisions.

Tax rates also affect commercial decisions, and the balance between individual and corporate taxes must reflect this. Accordingly, many countries have sought to attract new manufacturing industry with tax concessions. Finally, as rates rise, taxpayers seek more ways to avoid taxes. They employ tax advisers to find more tax-efficient routes, which, in particular, can involve a search for capital rather than income-yielding assets and the movement of activities overseas to less heavily taxed countries.

The balance between taxes. As the share of public expenditure in overall national income has risen, so has the strain on traditional sources of tax revenue. The original stalwarts, property and capital taxes, have shrunk in importance and been replaced by increasing reliance on income taxes, on social security contributions, and on sales taxes of various kinds. The balance between these taxes varies considerably among countries, which make differing decisions about the appropriate balance between taxes.

Each of the main types of tax is perceived by taxpayers in different ways. Social security taxes have everywhere risen in importance, partly as a result of the growth of social security expenditures but also because their association with the benefits received, however loose, reduces the unpopularity of increases.

Sales taxes are less obvious, as they change the price of goods rather than one's income. But sales taxes too have their limits; when the proportion of tax on a good is sufficiently high, consumption declines, and there is political pressure from consumers and industry to reduce the tax increases. Governments have been reluctant to increase indirect taxes significantly as the control of inflation has become a major policy goal.

Sale of goods and services. Taxation is not the only means by which a government can raise revenue. It can charge for the services it provides, or it can undertake profitable commercial activities. This is done to some degree by all Western governments, although the revenue raised is much less than that raised by taxation.

Other sources of revenue

Charging for public services faces a number of difficulties. Perhaps the most important is collection costs. Public services such as roads and parks are difficult to charge for, because they are closely integrated with the community. Some countries have tolls on major highways, and a few parks have admission charges, but in the main these are supplied free of charge. Other goods, such as museums and art galleries, are easier to charge for, but attempts at charging often generate more political opposition than can be justified by the limited revenue that could be raised.

A second consideration in deciding on charges is that it is rarely economically efficient to charge for public goods. Parks and roads, for example, have high initial costs but relatively low costs per user. Imposing a charge will mean that fewer people use them, and unless congestion is sufficiently severe to reduce others' enjoyment, less overall welfare will be generated.

In many countries many industries are publicly owned, and these include highly profitable industries such as those supplying gas or electricity. The profits from these industries provide revenue. It is an open question, however, whether they provide as much revenue as would the assets employed if these were invested in private sector companies. Other publicly owned undertakings make substantial losses, and this reduces the net value of any surplus.

(J.A.Ka.)

GOVERNMENT BORROWING

Although most of the resources required for public spending are raised each year through taxation, it is rare for any modern budget to balance in any one year. For a variety of reasons, ranging from a desire to accelerate capital spending to a policy of economic stabilization, governments may choose to raise some of their resources by borrowing rather than taxation. Many countries run an annual budget deficit, and the deficits have tended to increase in size. For some countries this means that the burden of the debt has been steadily increasing. In times of inflation, however, it may be possible for a government to run a deficit without actually increasing the real burden of debt, as inflation erodes the real value of its existing debt.

Forms of public debt. The necessity for governments to borrow in order to finance a deficit budget has led to the development of various forms of public debt, which are now a central feature of all capital markets. Governments may owe public debt in the form of bonds, notes, bills, and the like, which require specified payments to the holders at designated times. For the most part, public debt differs from private debt only in that it is an obligation of government rather than of private individuals or corporations.

Public versus private debt

Public debt may be classified according to various criteria.

External and internal debt. If the debt is held outside of the issuing jurisdiction, it is called external; if it is held within the jurisdiction, it is called internal.

Maturity period. Public debt ranges in maturity downward from infinity to periods of a month or even a few days. Debt instruments without a maturity date, requiring merely the payment of interest, are often called consols. The name originated in Great Britain, where the first important indeterminate-period debt issue happened to be one that consolidated a number of separate issues.

A large portion of government debt consists of bonds with specific maturities of five years to 99 years or more. Twenty- and 30-year periods are common. These are often known as long-term or funded debt.

Debt of maturity less than five years is often called short-term or floating debt and may take several forms: notes, with maturities from one to five years; treasury bills, with maturities from one month to a year and often sold at auc-

tion; and certificates of indebtedness, with similar maturity periods but available at a fixed interest rate.

Debt liquidity

The length of the maturity period affects what is known as the liquidity of the debt—*i.e.*, how quickly it can be converted into money. Securities with very short maturity periods are constantly repayable in money and thus have maximum liquidity. As the period of maturity increases, the liquidity falls, unless a capital loss is to be incurred, and the pure debt characteristic increases.

Type of issuer. Government debt may be directly issued by a government or by semiautonomous governmental organizations. Examples of the latter would include the railways and provincial power authorities in Canada and various federal lending agencies in the United States. Their issues may be guaranteed by the government (general obligation bonds) or may rest solely upon the enterprises themselves, to be paid out of their revenues. In the United States the latter type of obligation is known as a revenue bond.

Marketability. The great bulk of all government debt consists of marketable securities. These securities are negotiable and are sold freely on the market. They are usually issued in relatively large denominations, $1,000 or higher, and interest is paid by check or coupon on a periodic basis. Since they are salable, their price fluctuates from time to time, going above maturity value when the current market interest rate falls below the interest rate that they bear and falling below the maturity value when the current rate rises or when fear about the ability of the government to pay interest develops.

Other bonds (such as savings bonds) are not marketable but can be redeemed, at least after a specified period, for their principal plus accrued interest.

Other characteristics. Bondholders may receive current interest either by redemption of coupons attached to the bonds or by check from the government. Alternatively, interest may be receivable only upon maturity or redemption of the bond, as in the case of savings bonds. Interest and principal are usually payable in fixed monetary units, but they may be payable in amounts with fixed purchasing power based on changes in price levels.

Economics of government borrowing. Government borrowing is likely to have effects upon the economy substantially different from those of other methods of financing, and the existence of a sizable debt may likewise have important consequences. The effects of retiring (or repaying) the debt may also be significant. National government borrowing has the greatest impact, but that of subordinate units may have some influence as well.

Effects of borrowing. Government borrowing in the strict sense includes only borrowing from the private sector of the economy—from individuals, corporations, and various financial institutions, including banks. When the government obtains its funds from the central bank (the Bank of England, the Bank of Italy, the Bank of Japan, or the Federal Reserve System in the United States), it is really creating money rather than borrowing it, since the purchasing power is made by the central bank and no obligations to the public are created.

How governments borrow

When a government borrows, funds are transferred from the lender to the government, the lender exchanging his money for government securities. The effect is to reduce the liquidity of the lender—his command over cash—to an extent dependent upon the nature of the securities. The reduction in liquidity is small with short-term securities and greatest with nonsalable, nonredeemable securities—a type seldom issued except in time of war or other crises that create financial emergencies.

Funds loaned to the government almost certainly come from savings, unlike, for example, funds paid in higher taxes, which are more likely to come out of consumption. In many countries the major holders of public debt are, in fact, pension funds, which invest in government debt on behalf of the individual members of their pension schemes. To pay higher taxes, many individuals are forced to reduce their consumption since they have no margin of savings and are unable or unwilling to go into debt; others do so as a matter of choice, in an effort to keep their savings intact. Lending, on the other hand, is entirely voluntary. The person who buys government securities is not likely to increase his rate of saving or to decrease his consumption. If government borrowing raises the market rate of interest, this may in turn encourage the diversion of additional money to saving, as may government securities that offer additional attractions—such as small denominations or redeemability—not possessed by other securities. But both effects in total are not likely to be of any particular significance.

The net effect of government borrowing on total spending and thus on employment and national income depends upon its influence on real investment—the purchase of new capital goods. In a period of unemployment, when savings are available in greater quantity than is required for investment, government borrowing does not compete with private investment nor make it more costly. In effect, the government absorbs funds that would otherwise be idle.

Borrowing during full employment

In periods of full employment the situation is substantially different. With banks loaned up to the limit of their reserves and real investment absorbing all of savings, government borrowing will restrict private spending as much as an increase in taxation will under the same conditions.

Government borrowing is of economic significance in several other respects. First, the buying and selling of government securities provides the central bank with a means of influencing the money supply, essential for effective monetary policy. Second, borrowing avoids the adverse effects that taxes may have on incentives, particularly if the taxes are raised sharply above levels to which persons have become accustomed. Third, borrowing permits government expenditures to be higher than would otherwise be feasible. Finally, the foreign borrowing of some governments gives them access to a greater quantity of foreign exchange, which enables them to finance the import of capital goods essential for economic growth. This consideration is not of concern to highly developed countries.

Effects of debt. The existence of a government debt is of economic significance in itself, as distinct from the effects of the borrowing. In the first place, individuals who hold government securities regard them as a portion of their personal wealth. This is true even though the only way the government will ever pay the interest on the debt or repay the principal is by levying taxes on the community, which holds the debt. In this sense "we owe it to ourselves." But since these links are not immediately apparent, the existence of a debt may make individuals spend more on consumption and save less than they otherwise would. The additional consumption may reduce the rate of capital formation and economic growth; it may also increase the level of employment over what it would otherwise be.

Second, because government securities are more liquid than most other investments, their holders are able to increase consumption out of accumulated savings more easily than they could otherwise. This may contribute to inflationary pressures.

Third, if investors, and particularly the business community, regard the national debt as a source of potential economic instability, their willingness to undertake real investment will be lessened. At times, particularly in the 1930s, there has been widespread fear of government debt even though there was, in reality, little basis for the fear. A similar phenomenon sometimes arises in the case of subordinate units of government. A large debt may discourage expansion of economic activity because of the fear of high taxes in the future and the realization that the large debt may prevent borrowing for urgently needed local improvements.

Resistance to national debt

When governments borrow they must meet interest obligations, and these are usually paid out of taxes. The payment of interest on government debt thus involves a transfer of wealth from taxpayers to bondholders. The taxes may have adverse effects upon incentives, while receipt of the interest will provide no offset to these adverse effects. The tax-and-interest-payment program is also likely to redistribute wealth in favour of higher income groups, since government bonds are likely to be held to a greater extent by those groups. The effect may be to increase saving and reduce consumption.

Finally, large interest obligations lessen the ability of the

government to finance other governmental activities. This effect is particularly obvious at the local level, where there are limited tax potentials.

Retiring the debt. The retirement of government debt arising from a budget surplus has effects opposite from those of borrowing. Bondholders receive money in exchange for their bonds; though they could increase their consumption, they are more likely to put the funds into other securities and, as a consequence, security prices rise and money capital becomes more readily available for business investment. Whether it is used for that purpose depends, of course, on factors within the existing general economic situation.

Use of taxes for retiring debt

Money for retirement must be obtained from some source. If it is simply created, there is no repressive effect on consumption or investment, and total spending in the economy rises—although by an amount relatively small compared to the total retirement. If, as is more common, the debt is retired from tax revenues, consumption is reduced in substantial measure; the remainder of the tax is absorbed from savings, and real investment may be reduced. The net curtailment in spending from the program of debt retirement is likely to reduce total spending in the economy. Elimination of the debt has one other effect: while current taxes will be increased, future taxes required to meet interest and principal obligations will be reduced.

It is commonly thought that borrowing shifts the burden of governmental activities to future generations, since those generations will be assessed higher taxes to pay the interest and principal. Some economists have disputed the idea, noting that future generations will inherit both the bonds and the obligations to pay them and collectively will be neither richer nor poorer than if the debt had not been incurred, except as a result of the difficulties incident to the debt and its retirement noted in preceding sections. Regardless of the methods of financing, the real cost of any governmental activity, war or otherwise, is borne in the form of reduced private consumption and investment and harder work or the like during the period in which it is carried on. The only burden on the future is that arising from the depletion of natural resources, and this is not affected by methods of financing. Nevertheless, the method of financing may affect the way in which the burden of public expenditure is shared among different groups, including age groups.

Limitations on public sector debt. Although borrowing can often seem an attractive alternative to raising money from taxation or indeed to spending less, there are limits to how far a government can allow itself to become in debt either to its own citizens or to overseas investors (including intergovernmental agencies, such as the International Monetary Fund). Exceeding these limits can have serious results for the stability of the country concerned. When debt rises to unacceptable levels, so that investors cease to believe in the ability of the country's tax base to support it, then drastic measures are forced upon the country, including severe contraction of the economy.

Problems of borrowing. The desirability of government borrowing has been debated for centuries. The traditional argument against borrowing is, of course, the interest burden to which it gives rise, an argument applicable equally to private and governmental borrowing. These interest obligations require either higher levels of taxes, with possibly adverse effects on the economy, or reduced expenditures for other purposes. The payment of interest may easily result in a transfer of purchasing power to higher income groups, contrary to accepted standards of equity.

The interest burden

As well as causing more and more of the government's resources to be used to pay interest on its debt, a large public debt can push interest rates up for other borrowers. If the government is persuading a high proportion of available funds to be spent on public debt, the amount remaining for investment in other places, for example, investment in industry, is correspondingly small, with the result that a higher price (or higher interest rate) needs to be paid to attract such investment. This has been seen as a serious constraint in recent years, and most Western governments have tried to reduce their borrowing in order to keep interest rates down. There is some debate over just how important public borrowing is for interest rates, with monetarists believing it to be extremely important. Other types of economists are more skeptical, contending that factors such as inflation and the availability of private sector investment opportunities are more significant.

The financing of expenditures by borrowing instead of taxation and the debt itself, once incurred, increase total spending and so tend to produce higher prices and other inflationary effects in periods of full employment. During periods of full employment, any increase in government expenditures not offset by an equivalent decline in private spending for consumption or business expansion will be inflationary. This is the usual argument made against the use of borrowing instead of taxation from the standpoint of the goal of economic stability. It is primarily relevant to national government borrowing because the national government must assume the primary responsibility for lessening economic instability. But state and local borrowing is, of course, equally inflationary.

Borrowing, if freely employed, can easily lead to increases in government expenditures beyond levels regarded by society as the optimum and may reduce the pressures for efficiency and elimination of waste. As governments consider expenditure levels, the adverse reaction to taxation serves as an offset against the favourable response to increased services that will have to be paid for by taxation and thus facilitates the attainment of a balance between government-produced services and privately produced services. But if borrowing replaces taxation and is generally accepted as a suitable routine method of financing, the pendulum will swing in the direction of increased governmental activity, and appropriate balancing will be lost. The best evidence of this danger is to be found in the history of state and local government finance in the early 19th century in the United States, when large sums of money were borrowed for purposes of limited usefulness to society. Borrowing appears to be a less painful method of financing government, but, as has been noted, the costs of public expenditure still have to be met from current consumption or investment.

Restrictions on borrowing. Efforts have been made in some countries to set restrictions on government borrowing through legislative acts. In the United States, fear of excessive borrowing has resulted in restrictions on the amounts the executive, and even the legislative branches of government, can borrow. When many states found themselves in financial difficulties after borrowing heavily to provide funds for canals and railroads in the middle of the 19th century, public debt provisions were written into the constitutions of all but seven states. The provisions limiting borrowing differ widely. In most jurisdictions a maximum, usually expressed as an absolute dollar sum and one relatively low in terms of present-day expenditure levels, is set. Either this figure cannot be exceeded at all (except by amending the constitution) or it can be exceeded only with the approval of the voters at an election. In some places all bond issues require approval by popular vote and in some instances by more than a bare majority. The purposes for which funds may be borrowed and the duration of the issue are also frequently restricted. These constitutional restrictions have unquestionably lessened state borrowing; in so doing they have, perhaps, reduced waste, but they have also sometimes prevented urgently sought improvements. The limits have likewise greatly increased the use of revenue bonds, which are normally not subject to the restrictions. Unfortunately, the interest rate on these bonds is higher than the rate on other bonds.

Variations on limiting borrowing

Restrictions on municipal borrowing in the United States are almost universal. The restrictions, established either in the state constitutions or by state legislation, limit the total sum to be borrowed by any particular unit to a certain percentage (from 2 percent to more than 20 percent) of the total assessed value of its property. The limits vary for different types of local units (city, county, school district, etc.). They usually do not apply to debts incurred for self-liquidating enterprises. In many states every bond issue must be approved by popular vote, in some instances by a two-thirds majority. In other states the limits established may be exceeded by popular vote, often with a require-

U.S. state restrictions on borrowing

ment beyond a mere majority. Legislative controls also include maximum interest rates that may be paid, the duration of the issues, the purposes of the borrowing, and the establishment of means of retiring the bonds. Several states exercise review over local bond issues. Like the states, the local governments have found means of escaping the restrictions. Special taxing districts with their own debt limits are often formed when a city has reached its limit. Revenue bonds are also employed. In some states, such as Pennsylvania, there has been widespread creation of special authorities, school building authorities, for example, that have been established with the power to finance the building of schools by issuing revenue bonds. In turn, the authority pays interest and principal on the bonds from rentals obtained from the school districts using the buildings.

U.S. Congressional restrictions on borrowing

While there are no constitutional limits on federal borrowing powers in the United States, Congress for many years has restricted borrowing by the Treasury Department. Before 1917 borrowing was permitted only upon specific authorization by Congress. After 1917 maximum figures were set at first for each type of loan and then, after 1938, as an overall total. The 1938 figure of $45,000,000,000 was gradually increased to a high of $300,000,000,000 in 1945 and reduced to $275,000,000,000 in 1946. Buttressed by a strong belief prevailing in Congress that refusal to raise the limit would check growth in government spending, the limit remained at the 1946 level until 1954. Eventually, pressure on the limit became so great that various government bodies such as lending agencies were forced to borrow on their own at higher interest rates. A series of increases was made in the 1960s and 1970s, and by the early 1980s the limit exceeded $1,000,000,000,000. Experts differ in their estimates of the usefulness of the federal limit. Some believe that it curtails government waste and unjustified increases in expenditures, while others argue that it reduces flexibility in meeting emergencies, checks needed increases in various activities, could prevent quick action to stave off a depression, and leads to uneconomical forms of borrowing.

By the mid-1980s, the U.S. deficit approached an annual figure of $200,000,000,000 and was seen as a central economic problem. A movement grew for a constitutional amendment to prescribe a balanced federal budget. Such a constitutional provision would not, however, specify how such an outcome was to be achieved. Nor, given the many budgetary concepts described, would balance easily be defined. Congress instead passed the Gramm-Rudman-Hollings Act in 1985, which required arbitrary reductions in spending in all programs if the overall deficit failed to fall within certain limits that were set for the purpose of eliminating the deficit by the end of the decade.

Borrowing restrictions in Canada

In Canada, neither the dominion nor provincial governments are subject to debt limitations. Local government limits are comparable to those in the United States, and in several provinces bond issues must receive the approval of a provincial agency. In the United Kingdom borrowing by local governments is subject to control, and limits are usually established in terms of a ratio of debt to total ratable value (assessed value of property). After World War II much local borrowing was channeled through the Public Works Loan Board, and thus was subject to additional control. There are no arbitrary limitations on the amount the U.K. central government may borrow; effective limits are set by the reaction of capital markets and of interest rates to borrowing.

Wartime finance. The use of borrowing is regarded as inevitable in periods of major war. If taxes were increased sufficiently to finance all war costs, they could seriously impede the war effort by impairing incentives to work and by reducing the overall morale of the people. The limits of economically and politically tolerable taxation may well be below the maximum feasible allocation of resources to the war effort. Adequate tax increases would also aggravate the inequities of the tax structure; an overall level that would reduce total consumer spending to a level equal to the rate of output of consumer goods might well push some persons below subsistence levels and make it impossible for others to meet fixed commitments. While the use of borrowing as a method of war finance makes the control of inflation more difficult, there appears to be no escape from the necessity.

Evolution of government borrowing. The evolution of government borrowing was very slow. The extensive use of loans by governments became possible only after the ruler had become differentiated from the state and after the fact of the continuity of the state had been separated from the persons of the rulers. Other factors were also required: the development of a regular revenue source to provide funds for repayment of loans, a monetary system, and an organized money market. The earliest loans of medieval times were either forced loans or personal borrowing by the sovereign. Government borrowing in its modern form first occurred in medieval Genoa and Venice when the city governments borrowed on a commercial basis from the newly developed banks.

Early forms of government borrowing

In contemporary economies, the absolute figures of growth in government debt exaggerate the actual growth in the debt relative to the economy as a whole. In the first place, the general price level has increased significantly over recent decades; since debt obligations are stated in fixed monetary terms, the relative magnitude goes down as the price level goes up. The general rise in prices over a period thus reduces the problems created by the debt for the government and the magnitude of the adverse effects of the interest payments on the economy. The gain occurs at the expense of the bondholders, whose real economic position is worsened by the change.

Second, the rise in national income reflecting an increase in output reduces the real significance of a fixed sum of debt for the economy.

Sovereign debt. The oil crisis of 1973–74 and its aftermath created a new instability in world capital markets. Some countries, particularly Middle East producers with few economic activities not based on oil, gained revenues much in excess of their capacity to spend. Others, particularly in the less developed world, faced balance-of-payments problems that they found difficult to cover. Some other oil producers, such as Mexico, borrowed heavily in anticipation of rapidly increasing revenues. Those countries with surpluses of revenues over expenditures wanted to retain the liquidity of the financial assets that they acquired, and Western banks increasingly took on the role of intermediaries between the surplus and deficit countries. This led to the growth of sovereign lending—bank lending either to governments or to agencies of governments with government guarantees. While a bank lending to a private individual or company normally requires examination of the relationship of the loan to the borrower's assets, and of the interest to income or cash flow, banks felt able to apply more relaxed criteria to sovereign loans.

Complications of sovereign borrowing

By the early 1980s, however, it was apparent that for many countries sovereign debt had grown to levels at which even the interest on these loans would be met only by further borrowings. Moreover, these countries' limited capacity to repay might be undermined by political or economic instability. The problem was particularly acute in Latin America, where U.S. banks had lent aggressively. Argentina, Brazil, and Mexico had very large external debts; smaller countries such as Bolivia, Ecuador, and Peru had debt burdens that were even larger in relation to their capacity to service them. Similar difficulties were encountered in Africa and in parts of eastern Europe, particularly Poland.

The debtor countries were reluctant to repudiate their debts, which would have deprived them of access to the world capital markets and even perhaps to the world trading and payments system for a considerable time. At the same time, the lending banks were reluctant to demand repayment of their loans, which would have led to default and losses that would have wiped out a substantial portion of their reserves. Thus, there was a mutual interest in using the financial system to continue to support the indebted governments, and, paradoxically, the negotiating position of the borrowers was stronger than that of the lenders.

These were highly unstable arrangements, arousing fears that major defaults would occur. Such defaults might well set off a cumulative process of demands for repayment and

defaults that would undermine not only the economies of the debtor countries but also the banking and financial systems of the countries in which the lending institutions were located. Avoiding such a crisis has demanded continued sensitive responses on the part of international financial agencies, such as the International Monetary Fund and national regulatory institutions. (J.F.D./J.A.Ka.)

The economic functions of government

Over time, there have been considerable changes in emphasis on these different economic functions of the budget. In the 19th century, government finance was primarily concerned with the allocative function. The job of government was to raise revenue as cheaply and efficiently as possible to perform the limited tasks that it could do better than the private sector. A significant development in the intellectual history of the 20th century was the explicit recognition by economists, politicians, and the public at large of the importance of government in the operation of the economy. The national budget generally reflects the policy of the government toward the economy. Three main roles of the government in controlling the economy may be considered: the allocative function, the stabilization function, and the distributive function.

Government's main roles in controlling the economy

THE ALLOCATIVE FUNCTION

The allocative function in budgeting determines on what government revenue will be spent. Because a high proportion of national income is now devoted to public expenditure, allocation decisions become more significant in political and economic terms. In practice, most democracies contain a number of different factions that disagree on the proper allocation of resources and indeed the proper level of public sector involvement in the economy.

Public goods. Economists have sought to provide objective criteria for public expenditures through the so-called theory of public goods. It is generally recognized that some goods needed by the public cannot be provided through the private market. Lighthouses are a classic example. The costs of a lighthouse are such that no one shipowner will want to finance it; on the other hand, if a lighthouse is provided for one shipowner, it can be made available to all for no additional cost. Indeed it must be available to all, since there is no practical means of excluding ships from using the facility provided by the lighthouse, even if their owners have refused to pay for it. The only practical method of providing such services is by collective action.

Debate over paying for public works

If goods are to be provided in this way, rather than through the private market, it is immediately necessary to confront the twin problems of deciding how much to provide and who should pay for that provision. Even if all individuals wanted the service equally—as, perhaps, with lighthouses—their views on the extent of the service would be influenced by the allocation of the costs. Where different households may have different preferences and some may not want the service at all—as, for example, with defense by nuclear weapons—these difficulties are compounded. Economists have tried to devise abstract voting schemes that would reconcile these difficulties, but these appear to have little practical application.

Genuine public goods pose severe problems for the national budget; it is very difficult to decide how far particular goods—the arts, national parks, even defense—should be supplied, and therefore no formal procedure of determination is likely to evolve. What should be given to each will continue to be the subject of intense political debate, with allocation changing as the government changes. In the context of public policy, the efficient allocation of resources consists not merely of distributing funds in the pursuit of given objectives but also involves determining the objectives themselves.

Merit goods. The concept of merit goods assists governments in deciding which public or other goods should be supplied. Merit goods are commodities that the public sector provides free or cheaply because the government wishes to encourage their consumption. Goods such as subsidized housing or social services, which predominantly help the poor, or health care services, which help the poor and elderly, are generally regarded as having considerable merit and therefore have a strong claim on government resources. Other examples include the provision of retraining schemes or urban regeneration programs.

Cost-benefit analysis. Once decisions have been made on how the limited national budget should be divided between different groups of activities, or even before this, public authorities need to decide which specific projects should be undertaken. One method that has been used is cost-benefit analysis. This attempts to do for government programs what the forces of the marketplace do for business programs: to measure, and compare in terms of money, the discounted streams of future benefits and future costs associated with a proposed project. If the ratio of benefits to costs is considered satisfactory, the project should be undertaken. "Satisfactory" means, among other things, that the project is superior to any available public or private alternative. Or, if funds are limited, public investment projects may be assigned priorities according to their cost-benefit ratios.

One difficulty with cost-benefit analysis is that every government agency has an incentive to estimate favourable ratios for its own projects. It must, after all, compete with other agencies for funds. No one can be certain as to the returns to be expected from an irrigation canal or a highway. Private investors have also been known to exaggerate their claims in appealing to stockholders, but they are generally subject to market sanctions that encourage them to err on the side of caution.

In addition to the possibility that cost-benefit analysis may be biased by the preformed views of those commissioning the study, there are other, more fundamental difficulties. Almost all proposals have effects that are difficult to value in monetary terms. The siting of a new airport brings problems of noise and property blight to local people and increases the risk that civilians may die in an accident. Putting a sensible value on human life has been a continuing difficulty for those carrying out cost-benefit analyses, even though every project does in fact affect probabilities of life and death. These problems are, of course, not confined to cost-benefit analysis. Additional expenditure on health service or on road safety or better housing or heating old people's homes in winter all affect the number of people who die prematurely. The failure of cost-benefit analysis to provide answers to the problems of valuing life, or the quality of life, is a reflection of the wider problem confronting all decisions on public expenditure: the influence of subjective judgment.

Public ownership and privatization. Until the mid-1970s the proportion of economic activity controlled by the government and the share of taxes in national income tended to increase in most countries. Since then, however, challenges to this growth in the role of government have become increasingly influential, and moves to privatization have been common.

The trend to privatization

There are several types of privatization. One involves the sale to private owners of state-owned assets, and this is most correctly called privatization. Another is the sale of publicly owned industries, thus reversing the move to nationalization that occurred, particularly in western Europe, around and after World War II.

Privatization can also mean the dismantling of existing statutory restrictions on competition. State activities are often protected by legal prohibitions on competing private enterprise. The dismantling of such restrictions is seen as one method of improving the efficiency of state concerns.

Another aspect of privatization is the contracting out of publicly provided services. For example, U.S. municipalities entrust activities such as refuse collection, and in some cases even fire service, to private contractors.

Privatization and government revenue

While the objective of privatization is often to increase the efficiency of government activities, its implementation may also have important effects on government revenue. If an industry is sold for the present value of its expected earnings and if these earnings are the same in public and private ownership, privatization should have no net impact on public finances. If it is expected to be more efficient in the private sector, government finance, on balance, gains. If it is sold for less than the maximum revenue that would

be obtained—and this is often the case, either because of the difficulty of selling assets as large as nationalized industries or because the government wishes to secure a wide dispersion of share ownership—the impact is likely to be negative.

Other forms of government intervention. Government spending is not the only way in which government allocates resources. Its regional policies will determine whether domestic and overseas investors build factories in particular places, while its taxation policies will determine whether they build them at all. Government competition and merger policies affect the structure of industry and commerce, while regulatory activities—setting the number of hours shops may be open or who may buy cigarettes—have profound effects on commercial activities.

Government also affects allocations by setting the legal and administrative framework within which the economy functions. It may specify minimum wage levels or control the siting of new ventures and the activities of existing ones. Such activities of government profoundly affect the allocation of resources, but they are rarely monitored or subject to serious control.

THE STABILIZATION FUNCTION

Stabilization of the economy (*e.g.*, full employment, control of inflation, and an equitable balance of payments) is one of the goals that governments attempt to achieve through manipulation of fiscal and monetary policies. Fiscal policy relates to taxes and expenditures, monetary policy to financial markets and the supply of credit, money, and other financial assets.

History of stabilization policy. The use of fiscal and monetary policy as a means of stabilizing the economy is relatively recent, for the most part a development of the period after World War II. During the 19th century the only stabilization policy was that associated with the international gold standard. Under the gold standard, if a deficit occurred in a country's balance of payments, gold tended to flow out of the country. To counteract this process, the monetary authorities would raise interest rates and stiffen credit requirements, causing a fall in prices, income, and employment; this in turn led to a reduction in imports and an expansion of exports, thus improving the balance of payments. If a country had a surplus in its balance of payments, gold tended to flow in; this meant that interest rates fell and the supply of money and credit was increased. As a consequence, imports were stimulated and exports discouraged so that the surplus in the balance of payments tended to disappear. The adjustment mechanism also included another important element: capital movements between countries. When interest rates fell in surplus countries and rose in deficit countries, mobile international financial capital tended to flow from the former to the latter, contributing to the elimination of deficits and surpluses in the balance of payments. The working of this mechanism was partly automatic and partly the result of deliberate actions by the monetary authorities in each country.

The gold standard as a tool of stabilization

In this form of stabilization policy, external stability was achieved at the cost of stability in the domestic economy: fluctuations in domestic prices, incomes, and employment functioned as the levers for bringing about equilibrium in the balance of payments. Occasionally governments attempted to reduce the impact of this mechanism on the domestic economy, particularly on the price level. In particular, governments in some surplus countries took "sterilization actions" to prevent the gold inflow from increasing the supply of money and credit to the maximum extent. This could be done if the central bank offset its purchases of foreign exchange and gold with sales of government securities on the domestic credit market.

A somewhat more ambitious type of stabilization policy emerged in the period after World War I. During the 1920s unemployment in Great Britain rose to very high levels (between 20 and 30 percent of the labour force). Consequently, there was much discussion of whether employment could be increased by actions of the public authorities. Some maintained that budget deficits could raise the level of economic activity. An active part in this discussion was taken by British economist J.M. Keynes, and also by the Liberal Party, which in 1928 published proposals for government intervention entitled *Britain's Industrial Future*.

Stabilization policy of the 1920s and '30s

Stabilization theory. The new stabilization policy needed a theoretical rationale if it was ever to win general acceptance from the leaders of public opinion. The main credit for providing this belongs to Keynes. In his *General Theory of Employment, Interest and Money* (1935–36) he argued that high levels of unemployment might persist indefinitely unless governments took monetary and fiscal action. At that time he believed that fiscal action was likely to be more effective than monetary measures. Keynes's pessimistic view of monetary policy had a strong influence on economists and governments during and immediately after World War II.

The contribution of J.M. Keynes

Fiscal policy. Fiscal policy attempts to control the actions of individuals and companies by means of spending and taxation decisions. On the expenditure side, it can achieve this by spending money in ways—for example, on construction projects—that stimulate other activity, while on the taxation side it can affect work, investment, or production decisions by changing tax rates and levels. Fiscal policy thus has two major components: an overall effect generated by the balance between the resources the government puts into the economy through expenditures and the resources it takes out through taxation, charges, or borrowing; and a microeconomic effect generated by the specific policies it adopts. Both are important in stabilizing the economy—that is, they are countercyclical.

The two major components of fiscal policy

Overall fiscal policy involves the government in deciding whether it should spend more than it receives or less. Experience with countercyclical fiscal policy has been disappointing; in many cases, the lag between identifying the problem and fiscal response has been too long, with the result that a fiscal boost coincided with the next boom, while a contraction might coincide with the next recession. Fiscal policies that were intended to be countercyclical could end up exacerbating the original problems.

Another facet to fiscal policy is a government's attempt to guide the development of the economy by more specifically targeted policies. Thus most countries have from time to time attempted to cushion particular areas from the effects of a decline in their dominant industry by regional policies, to affect labour supply and demand by taxation, and to change the pattern of consumer purchases by changes to indirect taxes. These policies sometimes backfire as unforeseen consequences and interactions occur.

The heyday of fiscal stabilization policies occurred in the 1950s and '60s. In the 1970s governments became increasingly concerned about inflationary pressures, and important disturbances, particularly the oil crisis, disrupted world economies. Stabilization became a less important policy goal and one that governments were increasingly unable to achieve. Monetarist economic theories acquired increased influence until the turn of the 21st century, when some economists challenged the effectiveness of monetary policy and called for greater use of fiscal stabilization policies.

Monetary policy. Although the governmental budget is primarily concerned with fiscal policy (defining what resources it will raise and what it will spend), the government also has a number of tools that it can use to affect the economy through monetary control. By managing its portfolio of debt, it can affect interest rates, and by deciding on the amount of new money injected into the economy, it can affect the amount of cash in circulation and, therefore, indirectly affect prices and other economic variables.

Control of interest rates through debt management

At its simplest, monetarist theory postulates that in the economy there is a fixed amount of money, which circulates at a given velocity. This money is then available to finance the various transactions carried out in the economy at the prevailing prices. Under these circumstances, according to the theory, control of the price level can be maintained by controlling the amount of available money.

Although a desire to control inflation prompted many countries to adopt monetary policies, a number of different facets of economic behaviour can be affected by the use of monetary policy. In time of unemployment the central

bank may stimulate private investment expenditure, and possibly also household spending on consumer goods, by reducing interest rates and taking measures to increase the supply of credit, liquid assets, and money. The customary tools for doing this are open market operations, the discount rate of the central bank, and cash reserve requirements for commercial banks.

In open market operations the central bank buys government securities—bonds and treasury bills—from the private sector. The effect is to reduce interest rates by bidding up bond prices. In response, firms are likely to increase their investment expenditures, and households are likely to spend more on consumer goods.

The second tool of monetary policy, the discount rate of the central banks, is often used together with open market operations. If the discount rate is reduced, banks become more willing to extend credit to private borrowers because they can obtain funds themselves on easier terms. In many countries, changes in the discount rate tend to be followed by similar changes in the interest rates charged by banks to their borrowers.

The third tool of monetary policy, that of the cash reserve requirements (and, in some countries, certain types of government securities) for commercial banks, provides that banks must maintain money balances (in the form of deposits in the central bank) at a certain proportion of their liabilities. If the government reduces the reserve requirements, the banks can expand their loans further, thus increasing the total volume of credit outstanding.

Monetary policy, like fiscal policy, may also be used to combat inflationary tendencies by reversing the above measures; the central bank will then sell government securities (thereby increasing interest rates and reducing the supply of private credit and money), raise the discount rate, or increase reserve requirements.

Stabilization policy problems. Governments have displayed serious deficiencies in their ability to handle stabilization policy. Political leaders often lack economic information and understanding, and their economic advisers find it difficult to explain the economic situation to them and to apprise them of the relevant tools. There are also a variety of political inhibitions against taking action. One consequence is that what is designed to be a countercyclical policy becomes a procyclical one; instead of stabilizing the economy, it tends to destabilize it. The postwar experience in Britain is held by some to demonstrate the deficiencies of government in handling monetary and fiscal policy. In time of boom the government often followed an expansionary course; when a balance-of-payments crisis developed, it then took restrictive action—too late—and pushed the economy into deeper recession than would otherwise have occurred. On the basis of this experience, some economists have argued that a policy that did not attempt to counter the short-run swings in the economy would have been more successful in achieving stabilization. They maintain that the authorities should concentrate on letting the volume of money and credit increase steadily at a rate dictated by the long-term growth trend of the economy. Those who hold this view believe that capitalist economies are inherently stable, that crises are usually the result of bad policies on the part of the public authorities. Most economists do not share their optimism as to the stability of the economy if left alone; they continue to believe that governments must seek better tools for the purpose of short-run stabilization.

THE DISTRIBUTIVE FUNCTION

Virtually everything that a government does has some effect on the distribution of income or wealth at the various levels of society. Improvements in health care facilities benefit the sick, the old, and those about to have children. An increase in taxes on tobacco and beer affects the poor disproportionately, while an increase in capital taxes similarly affects the rich. Even regulatory and legislative activity benefits one group out of proportion to another. The redistributive consequences of the governmental budget can be reflected in a variety of ways; sometimes they are explicit and sometimes they are cited in the debate that follows the presentation of a budget. In the end, however, these consequences are hidden, unintended, and imperfectly understood. (A.Li./J.A.Ka.)

BIBLIOGRAPHY. Studies of the governmental budget include JESSE BURKHEAD, *Government Budgeting* (1956); E.S. KIRSCHEN *et al., Economic Policy in Our Time*, 3 vol. (1964); J.R. HICKS, *The Social Framework: An Introduction to Economics*, 4th ed. (1971); UNITED STATES CONGRESS, *Planning, Programming, Budgeting: Inquiry* (1970); DAVID J. OTT and ATTIAT F. OTT, *Federal Budget Policy*, 3rd ed. (1977); PETER C. SARANT, *Zero-Based Budgeting in the Public Sector: A Pragmatic Approach* (1978); and WILFRED BECKERMAN, *An Introduction to National Income Analysis*, 3rd ed. (1980). *Survey of Current Business* (monthly), published by the Bureau of Economic Analysis of the United States Department of Commerce, together with its supplement, *The National Income and Product Accounts of the United States: Statistical Tables* (irregular), provides practical analyses. The Brookings Institution annual *Setting National Priorities* explores U.S. budgetary developments; and the London Institute for Fiscal Studies does the same for the United Kingdom in its *IFS Report Series*.

Components of the budget and their interrelation are explored in RICHARD A. MUSGRAVE, *Fiscal Systems* (1969, reprinted 1981); HUGH HECLO and AARON WILDAVSKY, *The Private Government of Public Money*, 2nd ed. (1981); ALAN T. PEACOCK and JACK WISEMAN, *The Growth of Public Expenditure in the United Kingdom*, 2nd rev. ed. (1967); LEO PLIATZKY, *Getting and Spending: Public Expenditure, Employment, and Inflation*, rev. ed. (1984); CARL S. SHOUP, *Public Finance* (1969); HAROLD F. WILLIAMSON (ed.), *The Growth of the American Economy*, 2nd ed. (1951); ALAN S. BLINDER *et al., The Economics of Public Finance* (1974); CHARLES L. SCHULTZE, *The Politics and Economics of Public Spending* (1968); J.A. KAY and M.A. KING, *The British Tax System*, 3rd ed. (1983); CAROLYN WEBBER and AARON WILDAVSKY, *A History of Taxation and Expenditure in the Western World* (1986); MERVYN A. KING and DON FULLERTON (eds.), *The Taxation of Income from Capital: A Comparative Study in the United States, the United Kingdom, Sweden, and West Germany* (1984); JOHN F. DUE and JOHN L. MIKESELL, *Sales Taxation: State and Local Structure and Administration* (1983); CHARLES E. WALKER and MARK A. BLOOMFIELD (eds.), *New Directions in Federal Tax Policy for the 1980s* (1983); MARK ASHWORTH, JOHN HILLS, and NICK MORRIS, *Public Finances in Perspective* (1984); HENRY C. ADAMS, *Public Debts: An Essay in the Science of Finance* (1887, reprinted 1975); JOHN MAYNARD KEYNES, *How to Pay for the War* (1940); TILFORD C. GAINES, *Techniques of Treasury Debt Management* (1962); WARREN L. SMITH, *Debt Management in the United States* (1960); EDWARD NEVIN, *The Problem of the National Debt* (1954); HENRY C. MURPHY, *The National Debt in War and Transition* (1950); A. JAMES HEINS, *Constitutional Restrictions Against State Debt* (1963); JAMES M. BUCHANAN, *Public Principles of Public Debt* (1958); and JAMES M. FERGUSON (ed.), *Public Debt and Future Generations* (1964, reprinted 1982).

The economic role of the government is analyzed in C.F. BASTABLE, *Public Finance*, 3rd rev. ed. (1903, reprinted 1917); JAMES M. BUCHANAN and MARILYN R. FLOWERS, *The Public Finances*, 5th ed. (1980); RICHARD STONE and GIOVANNA STONE, *National Income and Expenditure*, 10th ed. (1977); BENT HANSEN, *Fiscal Policy in Seven Countries, 1955–1965: Belgium, France, Germany, Italy, Sweden, United Kingdom, United States* (1969); CHARLES J. HITCH and ROLAND N. MCKEAN, *The Economics of Defense in the Nuclear Age* (1960, reissued 1975); JOHN KENNETH GALBRAITH, *The Affluent Society*, 4th ed. (1984); FRANCIS M. BATOR, *The Question of Government Spending: Public Needs and Private Wants* (1960); LEIF JOHANSEN, *Public Economics* (1965, reissued 1975; originally published in Norwegian, 1965); ERIK LUNDBERG, *Instability and Economic Growth* (1968); RONALD L. TEIGEN (ed.), *Readings in Money: National Income, and Stabilization Policy*, 4th ed. (1978); C.A.E. GOODHART, *Money, Information, and Uncertainty* (1975); JOHN G. GURLEY and EDWARD S. SHAW, *Money in a Theory of Finance* (1960, reprinted 1971); VICTORIA CHICK, *The Theory of Monetary Policy*, rev. ed. (1977); JOSEPH A. PECHMAN and BENJAMIN A. OKNER, *Who Bears the Tax Burden?* (1974); and ROBERT H. HAVEMAN and JULIUS MARGOLIS (eds.), *Public Expenditure and Policy Analysis*, 3rd ed. (1983).

(K.E.P./J.F.D./A.Li./C.N.M./Ed.)

Graphic Design

Graphic design is the art and profession of selecting and arranging visual elements—such as typography, images, symbols, and colours—to convey a message to an audience. Sometimes it is called "visual communications," a term that emphasizes its function of giving form—*e.g.*, the design of a book, advertisement, logo, or Web site—to information. An important part of the designer's task is to combine visual and verbal elements into an ordered and effective whole. Graphic design is therefore a collaborative discipline: writers produce words and photographers and illustrators create images that the designer incorporates into a complete visual communication.

The evolution of graphic design as a practice and profession has been closely bound to technological innovations, societal needs, and the visual imagination of practitioners. Graphic design has been practiced in various forms throughout history; indeed, strong examples of graphic design date back to manuscripts in ancient China, Egypt, and Greece. As printing and book production developed in the 15th century, advances in graphic design developed alongside it over subsequent centuries, with compositors or typesetters often designing pages as they set the type.

In the late 19th century, graphic design emerged as a distinct profession in the West, in part because of the job specialization process that occurred there, and in part because of the new technologies and commercial possibilities brought about by the Industrial Revolution. New production methods led to the separation of the design of a communication medium (*e.g.*, a poster) from its actual production. Increasingly, over the course of the late 19th and the early 20th centuries, advertising agencies, book publishers, and magazines hired art directors who organized all visual elements of the communication and brought them into a harmonious whole, creating an expression appropriate to the content. In 1922 typographer William A. Dwiggins coined the term "graphic design" to identify the emerging field.

Throughout the 20th century, the technology available to designers continued to advance rapidly, as did the artistic and commercial possibilities for design. The profession expanded enormously, and graphic designers created, among other things, magazine pages, book jackets, posters, compact-disc covers, postage stamps, packaging, trademarks, signs, advertisements, kinetic titles for television programs and motion pictures, and Web sites. By the turn of the 21st century, graphic design had become a global profession, as advanced technology and industry spread throughout the world.

Typography is discussed in this essay as an element of the overall design of a visual communication. Similarly, the evolution of the printing process is discussed in this essay as it relates to developments in graphic design. For a complete history of these areas, see PRINTING, TYPOGRAPHY, and PHOTOENGRAVING.

Historical foundations 160
Manuscript design in antiquity and the Middle Ages 160
Early printing and graphic design 161
Graphic design in the 16th–18th centuries 161
Renaissance book design 161
Rococo graphic design 161
Neoclassical graphic design 162
Graphic design in the 19th century 163
The Industrial Revolution and design technology 163
William Morris and the private-press movement 163
Art Nouveau 164
Graphic design in the 20th century 164
Early developments 164
Modernist experiments between the world wars 165
Graphic design, 1945–75 166
The International Typographic Style
Postwar graphic design in the United States
Postwar graphic design in Japan
Graphic design, 1975–2000 167
Postmodern graphic design
Graphic design in developing nations
The digital revolution
Bibliography 168

Historical foundations

MANUSCRIPT DESIGN IN ANTIQUITY AND THE MIDDLE AGES

Although its advent as a profession is fairly recent, graphic design has roots that reach deep into antiquity. Illustrated manuscripts were made in ancient China, Egypt, Greece, and Rome. While early manuscript designers were not consciously creating "graphic designs," scribes and illustrators worked to create a blend of text and image that was at once harmonious and effective at conveying the idea of the manuscript. The ancient Egyptian Book of the Dead, which contained texts intended to aid the deceased in the afterlife, is a superb example of early graphic design. Hieroglyphic narratives penned by scribes are illustrated with colourful illustrations on rolls of papyrus. Words and pictures are unified into a cohesive whole: both elements are compressed into a horizontal band, the repetitive vertical structure of the writing is echoed in both the columns and the figures, and a consistent style of brushwork is used for the writing and drawing. Flat areas of colour are bound by firm brush contours that contrast vibrantly with the rich texture of the hieroglyphic writing.

Book of the Dead

During the Middle Ages, manuscript books preserved and propagated sacred writings. These early books were written and illustrated on sheets of treated animal skin called parchment, or vellum, and sewn together into a codex format with pages that turned like the pages of contemporary books. In Europe, monastic writing rooms had a clear division of labour that led to the design of books. A scholar versed in Greek and Latin headed the writing room and was responsible for the editorial content, design, and production of books. Scribes trained in lettering styles spent their days bent over writing tables, penning page after page of text. They indicated the place on page layouts where illustrations were to be added after the text was written, using a light sketch or a descriptive note jotted in the margin. Illuminators, or illustrators, rendered pictures and decorations in support of the text. In designing these works, monks were mindful of the educational value of pictures and the capacity of colour and ornament to create spiritual overtones.

Manuscript production in Europe during the Middle Ages generated a vast variety of page designs, illustration and lettering styles, and production techniques. Isolation and poor travel conditions allowed identifiable regional design styles to emerge. Some of the more distinctive medieval art and design approaches, including the Hiberno-Saxon style of Ireland and England and the International Gothic style prevalent in Europe in the late 14th and early 15th centuries, were used in manuscript books that achieved major graphic-design innovations. The Book of Kells, an illuminated gospel book believed to have been completed in the early 9th century at the Irish monastery of Kells, is renowned as one of the most beautiful Hiberno-Saxon manuscripts. Its page depicting the appearance of Christ's

Book of Kells

name in Matthew 1:18 is called the "Chi-rho page" (*c.* 800 AD). The design presents the monogram XPI—which was used to signify Christ in many manuscripts—as an intricately designed pattern of shimmering colour and spiraling forms blossoming over a whole page. The Book of Kells's Chi-rho page is a paradigm of how graphical form can become a metaphorical expression of spiritual experience: it clearly conveys the sacred nature of the religious content.

From the 10th through the 15th centuries, handmade manuscript books in Islāmic lands also achieved a masterful level of artistic and technical achievement, especially within the tradition of Persian miniature painting. The pinnacle of the Shīrāz school of Persian manuscript design and illustration is evident in a page by an unknown designer illustrating the great poet Nēẓāmī's *Khamseh* ("The Quintuplet"). This page from 1491 depicts the Persian king Khosrow II in front of the palace of his beloved, Shīrīn. Human figures, animals, buildings, and the landscape are presented as refined shapes defined by concise outlines. These two-dimensional planes are filled with vibrant colour and decorative patterns in a tightly interlocking composition. The calligraphic text, contained in a series of rectangles across the top and bottom of the page, balances the crisp illustrations.

EARLY PRINTING AND GRAPHIC DESIGN

While the creation of manuscripts led to such high points in graphic design, the art and practice of graphic design truly blossomed with the development of printmaking technologies such as movable type. Antecedents of these developments occurred in China, where the use of woodblock, or relief, printing, was developed perhaps as early as the 6th century AD. This process, which was accomplished by applying ink to a raised carved surface, allowed multiple copies of texts and images to be made quickly and economically. The Chinese also developed paper made from organic fibres by 105 AD. This paper provided an economical surface for writing or printing; other substrates, such as parchment and papyrus, were less plentiful and more costly to prepare than paper.

Early Chinese printing

Surviving artifacts show that the Chinese developed a wide range of uses for printing and that they achieved a high level of artistry in graphic design and printing from an early date. Artisans cut calligraphic symbols into woodblocks and printed them beautifully; printed sheets of paper bearing illustrations and religious texts were then pasted together to make printed scrolls. By the 9th or 10th century, paged woodblock books replaced scrolls, and literary, historical, and herbal works were published. Paper money and playing cards were also designed, the designs cut into woodblocks and printed. Chinese alchemist Pi Sheng invented a technique for printing with movable type about 1041–48 AD; however, this technology did not replace the hand-cut woodblock in Asia, in part because the hundreds of characters used in calligraphic languages made setting and filing the movable characters difficult.

Chinese inventions slowly spread across the Middle East and into Europe. By the 15th century, woodblock broadsides and books printed on paper were being made in Europe. By 1450 Johannes Gutenberg of Mainz, Germany, invented a method for printing text from raised alphabet characters cast on movable metal types. After this, printed books began to replace costly handmade manuscript books. Designers of early typographic books in Europe attempted to replicate manuscripts, often designing type styles based on current manuscript lettering styles. When the type was printed, spaces were left for illuminators to add pictures, ornate initials, and other decorative material. In this way, the compositor or typesetter was in effect the designer as he set the type. Some surviving copies of Gutenberg's landmark 42-line Bible have headers, initials, and sentence markers applied by hand in red and blue inks.

Gutenberg's Bible

Over time, typographic books developed their own design vocabulary. By the mid-15th century, printers combined woodblock illustrations with typeset text to create easily produced, illustrated printed books. They printed woodblock decorative borders and ornamental initials along with the type, subsequently having colour applied by hand to these printed elements. The first complete printed title page—identifying the book title, author, printer, and date—was designed in 1476.

The prevalence of movable type and increasingly advanced printing technology in Europe meant that, while other cultures continued to create manuscript designs and printed communications, major advances in graphic design over the next several centuries would often be centred in Europe.

Graphic design in the 16th–18th centuries

RENAISSANCE BOOK DESIGN

The Renaissance saw a revival, or "rebirth," of Classical learning from ancient Greece and Rome throughout Europe. Beginning in the late 15th century, printing played a major role in this process by making knowledge from the ancient world available to all readers. Typeface designs evolved toward what are now called Old Style types, which were inspired by capital letters found in ancient Roman inscriptions and by lowercase letters found in manuscript writing from the Carolingian period.

The Italian scholar and printer Aldus Manutius the Elder founded his Aldine Press in 1495 to produce printed editions of many Greek and Latin classics. His innovations included inexpensive, pocket-sized editions of books with cloth covers. About 1500 Manutius introduced the first italic typeface, cast from punches cut by type designer Francesco Griffo. Because more of these narrow letters that slanted to the right could be fit on a page, the new pocket-sized books could be set in fewer pages.

Aldine Press

The prototype for Renaissance book design was the Aldine Press's *Hypnerotomachia Poliphili* (1499; Figure 1), believed to be written by Francesco Colonna. The design of the work achieves an understated simplicity and tonal harmony, and its elegant synthesis of type and image has seldom been equaled. The layout combined exquisitely light woodcuts by an anonymous illustrator with roman types by Griffo utilizing new, smaller capitals; Griffo cut these types after careful study of Roman inscriptions. Importantly, double-page spreads were conceived in the book as unified designs, rather than as two separate pages.

During the 16th century, France became a centre for fine typography and book design. Geoffroy Tory—whose talents included design, engraving, and illustration, in addition to his work as a scholar and author—created books with types, ornaments, and illustrations that achieved the seemingly contradictory qualities of delicacy and complexity. In his 1531 *Book of Hours* he framed columns of roman type with modular borders; these exuberant forms were a perfect complement to his illustrations.

Geoffroy Tory

Typeface designer and punch-cutter Claude Garamond, one of Tory's pupils, achieved refinement and consistency in his Old Style fonts. Printers commissioned types from him rather than casting their own, making Garamond the first independent typefounder not directly associated with a printing firm. Works by Tory, Garamond, and many other graphic artists and printers created a standard of excellence in graphic design that spread beyond France.

The 17th century was a quiet time for graphic design. Apparently the stock of typeface designs, woodblock illustrations, and ornaments produced during the 16th century satisfied the needs of most printers, and additional innovation seemed unnecessary.

ROCOCO GRAPHIC DESIGN

The 18th-century Rococo movement, characterized by complex curvilinear decoration, found its graphic-design expression in the work of the French typefounder Pierre-Simon Fournier. After studying art and apprenticing at the Le Bé type foundry, Fournier opened his own type design and foundry operation. He pioneered standardized measurement through his table of proportions based on the French *pouce,* a now-obsolete unit of measure slightly longer than an inch. The resulting standard sizes of type enabled him to pioneer the "type family," a series of typefaces with differing stroke weights and letter widths whose similar sizes and design characteristics allowed them to be used together in an overall design. Fournier designed a

Pierre-Simon Fournier

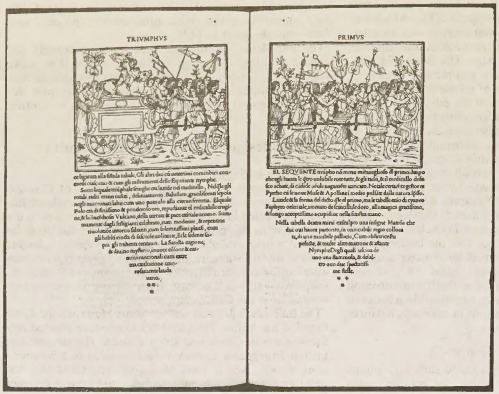


Figure 1: Page from the Aldine Press's *Hypnerotomachia Poliphili* (1499).
Library of Congress, Rosenwald Rare Book Collection

wide range of decorative ornaments and florid fonts, enabling French printers to create books with a decorative complexity that paralleled the architecture and interiors of the period. Because French law forbade typefounders from printing, Fournier often delivered made-up pages to the printer, thereby assuming the role of graphic designer.

Copperplate engraving became an important medium for book illustrations during this period. Lines were incised into a smooth metal plate, ink was pressed into these recessed lines, excess ink was wiped clean from the surface, and a sheet of paper was pressed onto the plate with sufficient pressure to transfer the ink from the printing plate to the paper. This allowed book illustrations to be produced with finer lines and greater detail than woodblock printing. In order to make text more compatible with these fine-line engravings, designers increasingly made casting types and ornaments with finer details. English engraver Robert Clee's engraved trading card (Figure 2) demonstrates the curvilinear decoration and fine detail achieved in both text and image by designers during the Rococo.

The Metropolitan Museum of Art, Gift of Mrs. Morris Hawkes, 1927.27.160(1)
All rights reserved. The Metropolitan Museum of Art



Figure 2: Black-and-white print of engraved trading card by Robert Clee, 18th century.

Graphic design often involves a collaboration of specialists. Eighteenth-century artists often specialized in book illustration. One such artist was Frenchman Charles Eisen, who illustrated French poet Jean de La Fontaine's *Contes et nouvelles en vers* (1762; *Tales and Novels in Verse*). In this work, Joseph Gerard Barbou, the printer, used types and ornaments by Fournier, full-page engravings by Eisen, and complex spot illustrations and tailpieces by Pierre-Phillippe Choffard. This superb example of Rococo book design combined the ornamented types, decorative initials, elaborate frames and rules, and intricate illustrations typical of the genre.

NEOCLASSICAL GRAPHIC DESIGN

In the second half of the 18th century, some designers tired of the Rococo style and instead sought inspiration from Classical art. This interest was inspired by recent archaeological finds, the popularity of travel in Greece, Italy, and Egypt, and the publication of information about Classical works. Neoclassical typographic designs used straight lines, rectilinear forms, and a restrained geometric ornamentation. John Baskerville, an English printer of the time, created book designs and typefaces that offered a transition between the Rococo and Neoclassical periods. His use of superbly designed types, printed on smooth paper without ornament or illustration, resulted in books of stately and restrained elegance. Baskerville's fonts had sharper serifs, more contrast between thick and thin strokes, and a more vertical, geometric axis, than Rococo typefaces.

John Baskerville

In the late decades of the 18th and early decades of the 19th century, Giambattista Bodoni, the Italian printer at the Royal Press (Stamperia Reale) of the Duke of Parma, achieved Neoclassical ideals in his books and typefaces. Bodoni laid forth his design statement in *Manuale tipografico* (1788; "Inventory of Types"); another edition of this book was published in 1818, after his death, by his widow and foreman. Bodoni advocated extraordinary pages for exceptional readers. He achieved a purity of form

with sparse pages, generous margins and line-spacing, and severe geometric types; this functional purity avoided any distractions from the act of reading. He drew inspiration from Baskerville as he evolved his preferences from Rococo-derived designs toward modern typefaces.

The Didot family

The Didot family of French printers, publishers, and typefounders also achieved Neoclassical ideals in their work. Books designed by the Didots have minimal decoration, generous margins, and simple linear borders. Pierre *l'aîné* Didot achieved technical perfection in his printing of the lavish *éditions du Louvre*. In these books, Pierre utilized types, designed at his brother Firmin's foundry, that provided a crisp counterpoint to the engraved illustrations by various artists working in the school of the French Neoclassical painter Jacques-Louis David. The illustrations of idealized figures in ancient Roman environments were engraved with flawless technique, obsessive detail, and sharp contrasts of light and shadow.

Graphic design in the 19th century

THE INDUSTRIAL REVOLUTION AND DESIGN TECHNOLOGY

The Industrial Revolution was a dynamic process that began in the late 18th century and lasted well into the 19th century. The agricultural and handicraft economies of the West had used human, animal, and water power, but they evolved into industrial manufacturing economies powered by steam engines, electricity, and internal-combustion motors. Many aspects of human activity were irrevocably changed. Society found new ways (often commercial) to use graphic designs and developed new technologies to produce them. Industrial technology lowered the cost of printing and paper, while making much larger pressruns possible, thus allowing a designer's work to reach a wider audience than ever before.

Posters

One popular medium for the graphic designer became the poster. Posters printed with large wood types were used extensively to advertise new modes of transportation, entertainment, and manufactured goods throughout the 19th century. This was possible in part because typefounders developed larger sizes of types for use on posted announcements and innovated new typefaces including sans serif, slab serif, and decorative designs. An American printer, Darius Wells, invented a lateral router that enabled the economical manufacture of abundant quantities of large wooden types, which cost less than half as much as large metal types. Wood-type posters usually had vertical formats; types of a mixture of sizes and styles were set in horizontal lines with a left-and-right alignment that created a visual unity.

The poster became even more popular as a result of advances in lithography, which had been invented about 1798 by Alois Senefelder of Bavaria. Building upon this discovery, colour lithographs, called chromolithographs, were widely used in the second half of the 19th century, and designers created increasingly colourful posters that filled the walls of cities. Designers of chromolithographic prints drew all the elements—text and image—as one piece of artwork; freed from the technical restraints of letterpress printing, they could invent fanciful ornaments and lettering styles at will. Many chromolithographs reflected an interest in the 1856 publication of English designer Owen Jones's *The Grammar of Ornament*, a methodical collection of design patterns and motifs that contained examples from Asian, African, and Western cultures. (Such explorations were consistent with the fascination with historicism and elaborate decoration found in architecture and product design during the Victorian era.)

Jules Chéret

Momentum for this poster-design approach began in France, where poster designer Jules Chéret was a pioneer of the movement. Beginning his career in 1867, he created large-scale lithographic posters that featured vibrant colours, textured areas juxtaposed against flat shapes, and happy, animated figures that captured *la belle epoch* of turn-of-the-century Paris (Plate 1). Chéret designed over one thousand posters during his career.

Chromolithography also made colourful pictures available to the homes of ordinary people for the first time in history. Designers developed ideas for packaged goods that were offered in tins printed with iconic images, bright colours, and embellished lettering. They also created trade cards, and "scrap," which were packets of printed images of birds, flowers, and other subjects collected by children.

As the century progressed, graphic design reached many people through magazines, newspapers, and books. The automation of typesetting, primarily through the Linotype machine, patented in the United States in 1884 by Ottmar Mergenthaler, made these media more readily available. One Linotype operator could do the work of seven or eight hand compositors, dramatically reducing the cost of typesetting and making printed matter less expensive.

WILLIAM MORRIS AND THE PRIVATE-PRESS MOVEMENT

During the 19th century, one by-product of industrialism was a decline in the quality of book design and production. Cheap, thin paper, shoddy presswork, drab, gray inks, and anemic text typefaces were often the order of the day. Near the end of the century, a book-design renaissance began as a direct result of the English Arts and Crafts Movement. William Morris, the leader of the movement, was a major figure in the evolution of design. Morris was actively involved in designing furniture, stained glass, textiles, wallpapers, and tapestries from the 1860s through the 1890s. Deeply concerned with the problems of industrialization and the factory system, Morris believed that a return to the craftsmanship and spiritual values of the Gothic period could restore balance to modern life. He rejected tasteless mass-produced goods and poor craftsmanship in favour of the beautiful, well-crafted objects he designed.

In 1888 Morris decided to establish a printing press to recapture the quality of books from the early decades of printing. His Kelmscott Press began to print books in 1891, using an old handpress, rich dense inks, and handmade paper. Decorative borders and initials designed by Morris and woodblocks of commissioned illustrations were cut by hand. Morris designed three typefaces based on types from the 1400s.

Kelmscott Press

The Kelmscott Press recaptured the beauty and high standards of incunabulum (texts produced when books were still copied by hand), and the book again became an art form. The press's masterwork is the ambitious 556-page *The Works of Geoffrey Chaucer*. Four years in the making, the Kelmscott Chaucer has 87 woodcut illustrations from drawings by renowned artist Edward Burne-Jones. For the single work, Morris designed 14 large borders, 18 smaller frames for the illustrations, and over 200 initial letters and words. An exhaustive effort was required by everyone involved in the project.

The influence of William Morris and the Kelmscott Press upon graphic design, particularly book design, was remarkable. Morris's concept of the well-designed page, his beautiful typefaces, and his sense of design unity—with the smallest detail relating to the total concept—inspired a new generation of graphic designers. His typographic pages, which formed the overwhelming majority of the pages in his books, were conceived and executed with readability in mind, another lesson heeded by younger designers. Morris's searching reexamination of earlier type styles and graphic-design history also touched off an energetic redesign process that resulted in a major improvement in the quality and variety of fonts available for design and printing; many designers directly imitated the style of the Kelmscott borders, initials, and type styles. More commercial areas of graphic design, such as job printing and advertising, were similarly revitalized by the success of Morris.

The Kelmscott Press's influence became immediately apparent in the rise of the private-press movement: printers and designers established small printing firms to design and print carefully crafted, limited-edition books of great beauty. Architect and designer Charles Robert Ashbee founded the Essex House Press in London, and bookbinder Thomas James Cobden-Sanderson joined printer Sir Emery Walker in establishing the Doves Press at Hammersmith. Books from the Doves Press, including its monumental masterpiece, the 1903 Doves Press Bible, are remarkably beautiful typographic books. They have no illustrations or ornaments; the press instead relied upon fine paper, perfect presswork, and exquisite type and spacing to

Doves Press Bible

produce inspired page designs. The Ashendene Press, directed by Englishman C.H. St. John Hornby, was another exceptional English private press. Following the example of Morris, these presses believed in the social value of making attractive and functional visual communications available to citizens of all walks of life.

In the United States, typeface designers, in particular Frederic W. Goudy and Morris F. Benton, revived traditional typefaces. Also inspired by the Arts and Crafts Movement, American book designer Bruce Rogers played a significant role in upgrading book design. By applying the ideals of the beautifully designed book to commercial production, Rogers set the standard for well-designed books in the early 20th century. An intuitive Classicist, Rogers possessed a fine sense of visual proportion. He also saw design as a decision-making process, feeling that subtle choices about margins, paper, type styles and sizes, and spatial position combine to create a unity and harmony. Type historian Beatrice Warde wrote that Rogers "managed to steal the Divine Fire which glowed in the Kelmscott Press books, and somehow be the first to bring it down to earth."

ART NOUVEAU

Art Nouveau was an international design movement that emerged and touched all of the design arts—architecture, fashion, furniture, graphic, and product design—during the 1890s and the early 20th century. Its defining characteristic was a sinuous curvilinear line. Art Nouveau graphic designs often utilized stylized abstract shapes, contoured lines, and flat space inspired by Japanese *ukiyo-e* woodblock prints. Artists in the West became aware of *ukiyo-e* prints as trade and communication between Eastern and Western nations increased during the last half of the 19th century. Building upon the example of the Japanese, Art Nouveau designers made colour, rather than tonal modeling, the primary visual attribute of their graphics.

Henri de Toulouse-Lautrec

One of the most innovative posters of the Art Nouveau movement was artist Henri de Toulouse-Lautrec's 1891 poster of the dancer La Goulue, who was performing at the Moulin Rouge. Toulouse-Lautrec captured the atmosphere and activity of the dance by reducing imagery to simple, flat shapes that convey an expression of the performance and environment. Although Toulouse-Lautrec produced only about three dozen posters, his early application of the *ukiyo-e* influence propelled graphic design toward more reductive imagery that signified, rather than depicted, the subject. He often integrated lettering with his imagery by drawing it in the same casual technique as the pictorial elements.

Alphonse Mucha, a young Czech artist who worked in Paris, is widely regarded as the graphic designer who took Art Nouveau to its ultimate visual expression. Beginning in the 1890s, he created designs—usually featuring beautiful young women whose hair and clothing swirl in rhythmic patterns—that achieved an idealized perfection. He organized into tight compositions lavish decorative elements inspired by Byzantine and Islāmic design, stylized lettering, and sinuous female forms. Like many other designers at the time, Mucha first captured public notice for poster designs, but he also received commissions for magazine covers, packages, book designs, publicity materials, and even postage stamps. In this way, the role and scope of graphic-design activity expanded throughout the period.

Will Bradley

Will Bradley, a self-taught American designer, emerged as another early practitioner of Art Nouveau. His magazine covers, lettering styles, and posters displayed a wide range of techniques and design approaches. Bradley synthesized inspiration from the European Art Nouveau and Arts and Crafts movements into a personal approach to visual imagery. By the 1890s, photoengraving processes (making printing plates from original artwork) had been perfected, which allowed reproductions to replicate the original artwork more accurately; hand engraving was often only an engraver's interpretation of an original. Bradley's work, in which he integrated words and picture into a dynamic whole, was printed from plates using this new technology.

Art Nouveau rejected historicism and emphasized formal invention, and so it became a transitional movement from Victorian design to the modern art movements of the early 20th century. This sense of transition is quite evident in the work of the Belgian artist and designer Henri van de Velde. After turning from Postimpressionist painting to furniture and graphic design in the 1890s, he used lines and shapes inspired by the natural world and abstracted them to the point that they appeared as "pure form"; that is, they appeared as abstract forms invented by the designer rather than as forms from nature. In works such as his poster for Tropon food concentrate (1899; Plate 1), undulating linear movements, organic shapes, and warm-hued colours combine into a nonobjective graphic expression. Although this poster has been interpreted as signifying the process of separating egg yolks and whites, the typical viewer perceives it as pure form.

Similarly exploring issues of form, and inspired in part by the theories and work of the American architect Frank Lloyd Wright, architects Charles Rennie Mackintosh and J. Herbert McNair joined artists (and sisters) Margaret and Frances Macdonald in a revolutionary period of creativity beginning in the 1890s. This group in Glasgow, Scotland, combined rectangular structure with romantic and religious imagery in their unorthodox furniture, crafts, and graphic designs. In a poster it made for the Glasgow Institute of Fine Arts (1895), for example, the group's emphasis upon rising vertical composition is evident.

Charles Rennie Mackintosh

Graphic design in the 20th century

EARLY DEVELOPMENTS

In the first decade of the 20th century, the experiments with pure form begun in the 1890s continued and evolved. Although the Glasgow group received a cool reception in the British Isles, designers in Austria and Germany were inspired by their move toward geometric structure and simplicity of form. In Austria, a group of young artists led by Gustav Klimt broke with the Künstlerhaus in 1897 and formed the Vienna Secession. These artists and architects rejected academic traditions and sought new modes of expression. In their exhibition posters and layouts for the Secession magazine, *Ver Sacrum,* members pushed graphic design in uncharted aesthetic directions. Koloman Moser's poster for the 13th Secession exhibition (1902) blends three figures, lettering, and geometric ornament into a modular whole. The work is composed of horizontal, vertical, and circular lines that define flat shapes of red, blue, and white. Moser and architect Josef Hoffmann were instrumental in establishing the Wiener Werkstätte ("Vienna Workshops"), which produced furniture and design objects.

Plakatstil

The German school of poster design called Plakatstil ("Poster Style") similarly continued the exploration of pure form. Initiated by Lucian Bernhard with his first poster in 1905, Plakatstil was characterized by a simple visual language of sign and shape. Designers reduced images of products to elemental, symbolic shapes placed over a flat background colour, and they lettered the product name in bold shapes. Plakatstil gained numerous adherents, including Hans Rudi Erdt, Julius Gipkens, and Julius Klinger.

Concurrent with these developments, in Germany Peter Behrens played an important role in graphic design. Behrens helped to develop a philosophy of Neue Sachlichkeit ("New Objectivity") in design, which emphasized technology, manufacturing processes, and function, with style subordinated to purpose. In 1907 Emil Rathenau, head of the AEG (Allgemeine Elektrizitäts Gesellschaft, a vast electrical manufacturing firm), appointed Behrens as artistic adviser for all of AEG's activities. Rathenau, a farsighted industrialist, believed industry needed the visual order and consistency that could only be provided by design. For AEG, Behrens developed what may be considered the first cohesive "visual identity system" (Plate 1); he consistently used the same logo, roman typeface styles, and geometric grids to create product catalogs, magazines, posters, other printed matter, and architectural graphics. Behrens's work for AEG was a harbinger of a major area of graphic design in the second half of the 20th century: the creation of a corporate identity through a program using trademarks, typefaces, formats, and colour in a consistent, controlled manner.

Corporate identity



Poster for a masked ball, designed by Jules Chéret, 1896.



Poster for Tropon food concentrate, designed by Henri van de Velde, 1899.



Logo of AEG (Allgemeine Elektrizitäts Gesellschaft), designed by Peter Behrens, 1907.



United States Army recruiting poster, designed by James Montgomery Flagg, 1917.

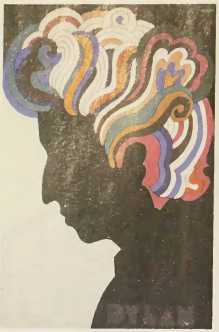


Spread from *Dlya golosa* (*For the Voice*) by Vladimir Vladimirovich Mayakovsky, designed by El Lissitzky, 1923.



Poster for the New York Subways Advertising Company, designed by Paul Rand, 1947.

Plate 1: (Top left, top centre, top right, bottom right) Collection of Philip B. Meggs. (centre) Library of Congress. (bottom left) Paul Rand.



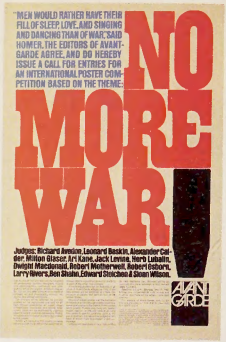
Poster of musician Bob Dylan, designed by Milton Glaser, 1967.



Spread from *Westvaco Inspirations 210*, designed by Bradbury Thompson, 1958.



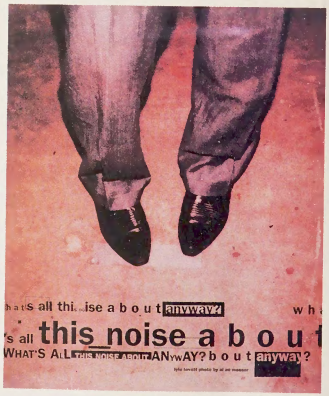
Poster commemorating the birth of the Iranian poet Saadi, designed by Ghobad Shiva, 1984.



Announcement for *Avant Garde* magazine's antiwar poster contest, designed by Herb Lubalin, 1968.

Poster proposal for Expo '85, designed by Igarashi Takenobu, 1982.





Page from *Beach Culture* magazine, designed by David Carson, 1990.

Plate 2: (Top left) © Milton Glaser, (top right) Westvaco Corporation, Westvaco Inspirations 210, 1958, by Bradbury Thompson, (centre) director and designer, Ghobad Shiva, client, UNESCO, (centre right) [illegible] design by David Carson, photography by Alan Messer, 1990, (bottom left) Takenobu Igarashi [illegible]

In addition to such aesthetic, commercial, and corporate purposes, graphic design also played an important political role in the early 20th century, as seen in posters and other graphic propaganda produced during World War I. Colour printing had advanced to a high level, and governments used poster designs to raise funds for the war effort, encourage productivity at home, present negative images of the enemy, encourage enlistment in the armed forces, and shore up citizens' morale. Plakatstil was used for many Axis posters, while the Allies primarily used magazine illustrators versed in realistic narrative images for their own propaganda posters. In American illustrator James Montgomery Flagg's recruiting poster (1917; Plate 1), for example, he used naturalistic portraiture of an allegorical person, appealing to the potential recruit, and bold lettering.

MODERNIST EXPERIMENTS BETWEEN THE WORLD WARS

Building upon the formal design experiments from the beginning of the century, between the world wars, European graphic designers utilized the new forms, organization of visual space, and expressive approaches to colour of such avant-garde movements as Cubism, Constructivism, De Stijl, Futurism, Suprematism, and Surrealism. Inspired by these movements, graphic designers pursued the most elemental forms of design. Such a concern with the essential formal elements of a medium characterize the Modernist experiments prevalent in all the arts of the period.

One pioneer of this approach was an American working in England, E. McKnight Kauffer, who was one of the first designers to understand how the elemental symbolic forms of Cubist and Futurist painting could be applied to the communicative medium of graphic design. Throughout the first half of the 20th century, his posters, book jackets, and other graphics achieved an immediacy and vitality well-suited to the fast-paced urban environment in which his visual communications were experienced.

Cassandre

Cassandre (the pseudonym of Adolphe Jean-Marie Mouron) used figurative geometry and modulated planes of colour, derived from Cubism, to revitalize postwar French poster design. From 1923 until 1936, Cassandre designed posters in which he reduced his subject matter to bold shapes and flat, modulated icons. He emphasized two-dimensional pattern, and he integrated lettering with his imagery to make a unified overall composition. Cassandre also utilized airbrushed blends and grading to soften rigid geometry. His clients included steamship lines and railway transportation, clothing, food, and beverage companies.

The austere visual language developed by artistic movements such as De Stijl in The Netherlands and by Suprematism and Constructivism in Russia influenced a Modernist approach to page layout. Suprematism, founded by Kasimir Malevich, inspired a young generation of designers to move toward a design based on the construction of simple geometric forms and elemental colour. Attributes of this approach in design included an underlying structure of geometric alignments, asymmetrical composition, elemental sans-serif typefaces, and simple geometric elements. Ornament was rejected, and open areas of white space were used as compositional elements. Works by the Russian Constructivist El Lissitzky exemplify this design approach. He developed design programs that utilized consistent type elements and placements. For example, his book design for Vladimir Vladimirovich Mayakovsky's *Dlya golosa* (*For the Voice*) (1923; Plate 1) is a seminal work of graphic design. The title spread for each poem is constructed into a dynamic visual composition, with geometric elements having symbolic meaning. In the title page to one poem, Lissitzky used a large red circle to signify the sun, the subject of the poem.

The Bauhaus

The Bauhaus, a German design school founded in 1919 with architect Walter Gropius as its director, became a crucible where the myriad ideas of modern art movements were examined and synthesized into a cohesive design movement. In its initial years, the Bauhaus held an Expressionist and utopian view of design, but it later moved toward a functionalist approach. Bauhaus artists and designers sought to achieve a unity between art and technology and to create functional designs—often utilizing the pure forms of Modernism—that expressed the mechanization of the machine age. In 1923 the Hungarian Constructivist László Moholy-Nagy joined the faculty. Among his numerous contributions, Moholy-Nagy introduced a theoretical approach to visual communications. Important in his theory was the use of photomontage (a composite photographic image made by pasting together or superimposing different elements) as an illustrative medium. He promoted the integration of words and images into one unified composition and the use of functional typography.

Herbert Bayer was appointed first master of the newly founded Druck and Reklame ("Printing and Advertising") workshop at the Bauhaus in 1925. Bayer's poster for Wassily Kandinsky's 60th-birthday exhibition (1926) incorporates Constructivist and De Stijl influences. It clearly embodies the Bauhaus design philosophy: elemental forms are shorn of ornament, and forms are selected and arranged in order to serve a functional purpose ("clarity of information"), with a visual hierarchy of size and placement in descending prominence from the most important to secondary facts. The elements are masterfully balanced and aligned to create a cohesive composition, and the tilting at a diagonal angle energizes the space.

The unprecedented graphic designs produced during this period were explained and demonstrated to printers and designers through writings and designs by Jan Tschichold, a young German designer. As a result, many designers in Europe and throughout the world embraced this new approach to graphic design. An announcement for Tschichold's book *Die Neue Typographie* (1928; "The New Typography") typifies his own philosophy. Tschichold advocated functional design that uses the most direct means possible. His systematic methodology emphasized contrast of type sizes, widths, and weights, and he used white space and spatial intervals as design elements to separate and organize material. He included only elements that were essential to the content and page structure.

Jan Tschichold

Many designers sought other ways to use geometry to evoke a modern spirit for the machine age. *Art Deco*, *streamline*, and *moderne* are terms used to denote the loosely defined trend in art, architecture, and design from the 1920s to the 1940s that utilized decorative, geometric designs. Everything from skyscrapers to furniture to—in the case of graphic design—cosmetics packaging, posters, and typefaces used zigzag forms, sunbursts, and sleek geometric lines to project a feeling of a new technological era.

At the same time, a number of Dutch designers, including Piet Zwart, drew upon the Modernist vocabulary of form and colour to develop unique personal approaches to graphic design, applying their vision to the needs of clients. While working at an architectural firm in the early 1920s, Zwart received commissions for graphic-design projects by happenstance. In his work from the 1920s and '30s, he rejected the conventional norms of typography and instead approached the layout of an ad or brochure as a spatial field upon which he created dynamic movements and arresting forms. This is seen, for example, in a dynamic advertisement for Nederlandsche Kabelfabriek Delft (NKF) cable factory (1924), which proclaims, "Normaal cable [is the] best cable for the price." Zwart believed the fast pace of 20th-century life meant viewers had little time for lengthy advertising copy. He used brief telegraphic text, bold typefaces placed at an angle, and bright colours to attract attention and to convey his client's message quickly and effectively.

Swiss poster design

Swiss designers also brought tremendous vitality to graphic design during this period. After studying in Paris with Fernand Léger and assisting Cassandre on poster projects, Herbert Matter returned to his native Switzerland, where from 1932 to 1936 he designed posters for the Swiss Tourist Board, using his own photographs as source material. He employed the techniques of photomontage and collage in his posters, as well as dynamic scale changes, large close-up images, extreme high and low viewpoints, and very tight cropping of images. Matter carefully integrated type and photographs into a total design.

When the Nazis rose to power in Europe during the 1930s, Modernist experiments were denounced, and many artists, architects, and designers immigrated to the United

States. This migration, along with their professional and teaching activities, would play a major role in shaping post-war American art and design. During World War II, posters were used once again as a major form of political propaganda, although they functioned alongside radio broadcasts and propaganda films in governmental war efforts.

GRAPHIC DESIGN, 1945–75

The International Typographic Style. After World War II, designers in Switzerland and Germany codified Modernist graphic design into a cohesive movement called Swiss Design, or the International Typographic Style. These designers sought a neutral and objective approach that emphasized rational planning and de-emphasized the subjective, or individual, expression. They constructed modular grids of horizontal and vertical lines and used them as a structure to regularize and align the elements in their designs. These designers preferred photography (another technical advance that drove the development of graphic design) as a source for imagery because of its machine-made precision and its ability to make an unbiased record of the subject. They created asymmetrical layouts, and they embraced the prewar designers' preference for sans-serif typefaces. The elemental forms of the style possessed harmony and clarity, and adherents considered these forms to be an appropriate expression of the postwar scientific and technological age.

Josef Müller-Brockmann

Josef Müller-Brockmann was a leading designer, educator, and writer who helped define this style. His poster, publication, and advertising designs are paradigms of the movement. In a long series of Zürich concert posters, Müller-Brockmann used colour, an arrangement of elemental geometric forms, and type to express the structural and rhythmic qualities of music. A 1955 poster for music by Igor Stravinsky, Wolfgang Fortner, and Alban Berg demonstrates these properties, along with Müller-Brockmann's belief that using one typeface in two sizes (display and text) makes the message clear to the audience.

The programmatic uniformity of this movement would be widely adopted by designers working in the area of visual identity systems during the second half of the 20th century. Multinational corporations soon adopted the tenets of the International Typographic Style: namely, the standardized use of trademarks, colours, and typefaces; the use of consistent grid formats for signs and publications; the preference for the contemporary ambiance of sans-serif types; and the banishment of ornament.

Postwar graphic design in the United States. While designers in Europe were forging the International Typographic Style into a cohesive movement, American designers were synthesizing concepts from modern art into highly individualistic and expressive visual statements. From the 1940s through the 1960s, New York City was a major centre for innovation in design and the fine arts.

Paul Rand

During the 1940s, Paul Rand emerged as an American designer with a personal and innovative approach to modern design. Rand understood the vitality and symbolic power of colour and shape in the work of artists such as Paul Klee, Wassily Kandinsky, and Pablo Picasso. In a 1947 poster promoting New York subway advertising (Plate 1), for example, Rand created a design from elemental geometric forms and colours that can be read as both an abstracted figure as well as a target, conveying the concept that one can "hit the bull's-eye," or reach potential audiences for plays, stores, and other goods and services by advertising in the subway. An ordinary message is rendered extraordinary through the power of visual forms and symbols. Rand's work spanned a range of graphic media including advertising, book jackets, children's books, corporate literature (such as annual reports), packaging, posters, trademarks, and typefaces.

In the 1950s Rand began to spend more of his time on corporate image projects, and he designed what would become ubiquitous trademarks and visual identities for major corporations including IBM, Westinghouse, ABC television, and UPS. Many other prominent designers—including Saul Bass (whose many visual identity programs included logos for AT&T), Lester Beall, and the partnership of Tom Geismar and Ivan Chermayeff—focused their practices upon corporate design, as multinational corporations understood the need for consistent graphic standards in their facilities and communications.

Bradbury Thompson, a prominent magazine art director, designed a publication called *Westvaco Inspirations* for a major paper manufacturer from 1938 until the early 1960s. His playful and innovative approach to type and imagery is shown in the design of a spread from *Westvaco Inspirations 210* (1958; Plate 2). Here, Thompson responded to the geometric forms of African masks in the Ben Somoroff photograph in the spread by "drawing" a masklike face out of letters spelling "Westvaco." Thompson's complex layouts combined art with coloured shapes and unusual typographic arrangements. He explored printing techniques by separating the four plates used to print full-colour images—cyan (a warm blue), magenta, yellow, and black—and having them printed in different positions on the page. He also had engravings from old books enlarged and overprinted in unexpected colours. These experiments were very influential, as they showed a generation of designers new possibilities.

Magazine design

Magazines placed more emphasis upon graphic design during the postwar period. Alexey Brodovitch, the art director of *Harper's Bazaar* from 1934 until 1958, pioneered a new approach to magazine design. He created a flowing perceptual experience for the reader who paged through his magazines by varying sizes of type and imagery, alternating complex pages with simple layouts containing large areas of white space, and creating an overall sense of rhythmic movement. The beauty of Brodovitch's designs was enhanced by the impressive team of collaborators at *Bazaar*, which included photographer Richard Avedon.

The postwar period has been called a "golden age" of magazine design, when art directors including Henry Wolf (at *Esquire* and *Harper's Bazaar*) and Otto Storch (at *McCall's*) extended Brodovitch's imaginative approach to page layout in large-format magazines. Storch believed concept, text, type, and image should be inseparable in editorial design, and he applied this belief to the editorial pages of *McCall's*.

Motion graphics

The emergence of television began to alter the roles of print media and graphic design, while also creating new opportunities for designers to work on television commercials and on-air graphics. "Motion graphics" are kinetic graphic designs for film titles and television that occur in the fourth dimension—time. A variety of animated film techniques were applied to motion-picture titling in the 1950s by Saul Bass and, in Canada, by Norman McLaren of the Canadian National Film Board. For example, Bass's titles for Otto Preminger's 1959 film *Anatomy of a Murder* reduce a prone figure to disjointed parts, which move onto the screen in carefully orchestrated sequences that conclude with their positioning to form the figure; the lettering of the film's title appears as part of the sequence.

Vernacular imagery and popular culture inspired a generation of American designer/illustrators who began their careers after World War II, including the 1954 founders of the Push Pin Studio in New York. Their work combined a fascination with the graphic simplicity and directness of comic books with a sophisticated understanding of modern art, especially of Surrealism and Cubism. The Push Pin artists' unabashedly eclectic interest in art and design history led them to incorporate influences ranging from Persian rugs to children's art and decorative Victorian typefaces. In their work, a graphic vibrancy supported a strong conceptual approach to the visual message.

Political design

Several major directions emerged in American graphic design in the 1960s. Political and social upheavals of the decade were accompanied by a resurgence of poster art addressing the Civil Rights Movement, the women's movement, environmentalism, and the Vietnam War. Placing ads on radio and television was beyond the economic means of most private citizens, independent art groups, and social-activist organizations; however, they could afford to print and distribute flyers and posters, and they could even sell their posters to public sympathizers to raise money for their causes.

As popular music became increasingly culturally significant, graphics for the recording industry emerged as a locus

of design creativity. One Push Pin Studio founder, Milton Glaser, captured the imagination of a generation with his stylized curvilinear drawing, bold flat colour, and original concepts. Glaser's poster (1967; Plate 2) for folk-rock musician Bob Dylan is one of many music graphics from the 1960s that achieved an iconic presence not unlike that of Flagg's "I Want You" poster from World War I. Over the course of the second half of the century, Glaser steadily expanded his interests to include magazine design, restaurant and retail store interiors, and visual identity systems.

The 1960s also saw the rapid decline of hand- and machine-set metal type, as they were replaced by display-and-keyboard phototype systems. Since it is very inexpensive to produce new typefaces for photographic typesetting, the widespread use of phototype systems set off a spate of new designs and reissues of long-unavailable typefaces, such as decorative Victorian wood types. American Herb Lubalin is notable among the designers who embraced the new flexibility phototype made possible for designers. Type could be set in any size, the spaces between letters and lines could be compressed, and letters could be expanded, condensed, touched, overlapped, or slanted. Lubalin's ability to make powerful visual communications solely with type is seen in a 1968 announcement (Plate 2) for an antiwar poster contest sponsored by Avant Garde magazine. The magazine's logo, placed in the dot of the exclamation point, uses ligatures (two or more letters combined into one form) and alternate characters to form a tightly compressed image. This logo was developed into a typeface named Avant Garde, one of the most successful and widely used fonts of the phototype period.

Advertising

A creative revolution in advertising writing and design also occurred during this period. Advertising agencies approached marketing objectives through the use of witty headlines, simple layouts, and clever visual images. Copywriters and art directors, working as collaborative creative teams, sought a synergy between word and image. The Doyle Dane Bernbach advertising agency played an influential role in the history of graphic design by creating advertisements that spoke intelligently to consumers and avoided the hyperbole of the typical "hard sell."

One of the many advertising designers who launched his career at Doyle Dane Bernbach was George Lois, whose works were engagingly simple and direct. Lois went on to design over 90 covers for *Esquire* magazine in the 1960s. He used powerful photographs and photomontages, usually by Carl Fischer, to make succinct editorial statements about the United States. These designs acted as independent visual/verbal statements about such topics as assassinations and civil rights.

Postwar graphic design in Japan. During the 1960s and '70s, American graphics from the New York area, as well as European graphics from the International Typographic Style, influenced designers around the world.

In postwar Japan, for example, when the country emerged as a major industrial power, graphic design evolved into a major profession serving the needs of industry and cultural institutions. European Constructivism and Western design exerted an important influence on Japanese design, but these lessons were assimilated with traditional Japanese art theory. For example, the Japanese tradition of family crests inspired many Japanese designers' approach to trademark design. Similarly, symmetrical composition, central placement of iconic forms, harmonious colour palettes, and meticulous craftsmanship—all characteristics of much of Japanese art—were often elements of Japanese graphics.

Kamekura Yūsaku

The first generation of graphic designers to emerge after the war was led by Kamekura Yūsaku, whose importance to the emerging graphic-design community led to the affectionate nickname "Boss." Kamekura's poster proposal (1967) for the Japanese World Expo '70 in Ōsaka, for example, displays his ability to combine 20th-century Modernist formal experiments with a traditional Japanese sense of harmony.

In counterpoint to the formalist tendencies found in much Japanese graphic design, some Japanese designers drew upon other sources of inspiration to arrive at individual approaches to visual-communications problems. Iconography from diverse mass media—including comic books, popular science-fiction movies, and newspaper photographs—provided a rich vocabulary for Yokoo Tadanori, whose work beginning in the 1960s inspired a new generation of Japanese designers. In his early posters and magazine covers he utilized a variety of contemporary techniques; for example, he used crisp line drawings to contain photomechanical screens of colour. He worked in a Pop-art idiom, but he used revered Japanese imagery as source material, rather than the contemporary imagery usually found in Pop art. In his poster publicizing four nō theatre productions (1969), for example, he placed iconic images on a luminous gold-and-blue field, combining traditional imagery with a contemporary sense of whimsy. Over time, montage effects became increasingly important to Yokoo, as he built his designs from photographic and graphic elements filled with dramatic luminosity.

A very different vision emerged in the work of Satō Kōichi, who from the 1970s created an otherworldly, metaphysical design statement. He used softly glowing blends of colour, richly coloured and modulated calligraphy, and stylized illustrations to create poetic visual statements that ranged from contemplative quietude to celebratory exuberance. For example, in his poster for a musical play—which was itself adapted from a nursery rhyme about soap bubbles—Satō combined an astronomical sky chart and a handprint glowing with a lavender-and-blue aura to evoke a feeling of ephemeral atmospheric space. Such designs achieve a rare level of visual poetry.

GRAPHIC DESIGN, 1975–2000

Postmodern graphic design. By the late 1970s, many international architectural, product, and graphic designers working in the Modernist tradition thought that the movement had become academic and lost its capacity for innovation. Younger designers challenged and rejected the tenets of Modernism and questioned the "form-follows-function" philosophy that came to be associated with the diluted, corporate version of Modernism that derived from the International Typographic Style. Designers began to establish and then violate grid patterns; to invert expected forms; to explore historical and decorative elements; and to inject subjective—even eccentric—concepts into design. This reaction to Modernist developments is called postmodernism, and it took design in many new directions.

Women and design

During the late 1970s, April Greiman was acclaimed for her postmodernist experimentation. (In the 1970s and '80s, increasing numbers of women entered the graphic-design field and achieved prominence.) Her dynamic typographic innovations and colourful montages were often made in collaboration with photographer Jayme Odgers. A cover for *Wet* magazine, for example, evokes the vibrant cultural scene in southern California. In this work from 1979, a colour photocopy of singer Ricky Nelson, collaged images from magazines, Japanese papers, and airbrushed blends of colour are combined into a cohesive design. Greiman also explored the application of video imagery to print graphics.

The dynamic spatial arrangement and decorative geometric patterns that enliven many postmodern designs are seen in a 1983 poster designed by William Longhauser. The letters forming the last name of postmodern architect Michael Graves become fanciful edifices, which echo the patterns and textures found in Graves's buildings. As with much postmodern design, the result is strikingly original.

Such a disruption of expected forms and grids was also apparent in the work of Japanese designer Igarashi Takenobu. After studying design fundamentals in Los Angeles, Igarashi began his independent design practice in Tokyo and used basic design elements—point, line, plane, grids, and isometric perspectives—as the building blocks of his work. This design vocabulary enabled him to invent imaginative solutions. His poster proposal (1982; Plate 2) for Expo '85, an international exposition of the dwelling and construction industry, turns the letters into structural forms pulled apart to reveal their inner structures. His experimentation with form fulfilled both an aesthetic and a commercial purpose: the deconstructed forms clearly make reference to his client, the construction industry.

Graphic design in developing nations. Late in the 20th century, increasingly accomplished graphic-design activity began to appear in developing nations. These advancements occurred because of a number of factors, including expanded access to professional education at local schools and abroad, the increased availability of computer and printing technology, and a growing base of industrial, cultural, and communications-industry clients. Designers from these nations often drew upon established design approaches from industrialized nations, but they commingled these lessons with local and national traditions in their quest for effective visual communications.

Design in the Middle East

In the Middle East, graphic designers often applied new technology to depictions of traditional subject matter and iconography. Throughout the late 20th century, Iranian graphic designer Ghobad Shiva evoked the colour palette, traditional Arabic calligraphy, and page layouts of ancient Persian manuscripts in his graphic work, which ranged from packaging to advertising and editorial design to stage sets. His poster (1984; Plate 2) celebrating the 800th anniversary of the birth of the Iranian poet Saadi, for example, displays his exquisite control of colour and his ability to create a vibrant image. These stylized illustrations continued the traditions of ancient Persian manuscripts, but within the context of a contemporary design idiom.

Design in Africa

Graphic design developed slowly in Africa after World War II, but by the end of the 20th century, a number of designers there received international acclaim for their individual creations. In Zimbabwe, filmmaker and designer Chaz Maviyane-Davies created films and graphic designs in the late 1980s and the 1990s. His posters, advertising designs, and magazine covers captured the spirit and life of his nation and often promoted social change. His interest in photographic and cinematic symbolism is apparent in a powerful library poster in which a young man has wings that are covered with printed matter from the library. Here, the image and social message—"Knowledge will set you free"—unite in a powerful visual statement.

Design in Latin America

In Latin America, professional graphic design similarly developed slowly after World War II. Eventually, in Argentina, and then in other nations, a graphic-design profession began to evolve. Latin American designers often built upon European and North American influences to develop distinctive communication designs. For example, a film festival poster (1992) by Venezuelan designer Santiago Pol utilizes clear symbolic forms within a highly sophisticated spatial configuration, both elements of Modernist graphic design. In this work, dynamic shapes signify three peppers, symbols that are redolent with regional symbolism; the central pepper is formed by the white, or negative, space between the red and green ones. These peppers are punctuated by film sprocket holes, which connect the image to the poster's theme of film. In this way, Pol's creative combination of symbols provides a distinct regional image for the film festival.

The digital revolution. Until the late 20th century, the graphic-design discipline had been based on handicraft processes: layouts were drawn by hand in order to visualize a design; type was specified and ordered from a typesetter; and type proofs and photostats of images were assembled in position on heavy paper or board for photographic reproduction and platemaking. Over the course of the 1980s and early '90s, however, rapid advances in digital computer hardware and software radically altered graphic design.

Computer graphics

Software for Apple Computer's 1984 Macintosh computer, such as the MacPaint™ program by programmer Bill Adkinson and graphic designer Susan Kare, had a revolutionary human interface. Tool icons controlled by a mouse or graphics tablet enabled designers and artists to use computer graphics in an intuitive manner. The Postscript™ page-description language from Adobe Systems, Inc., enabled pages of type and images to be assembled into graphic designs on screen. By the mid-1990s, the transition of graphic design from a drafting-table activity to an on-screen computer activity was virtually complete.

Digital computers placed typesetting tools into the hands of individual designers, and so a period of experimentation occurred in the design of new and unusual typefaces and page layouts. Type and images were layered, fragmented, and dismembered; type columns were overlapped and run at very long or short line lengths; and the sizes, weights, and typefaces were often changed within single headlines, columns, and words. Much of this research took place in design education at art schools and universities. American designer David Carson—art director of *Beach Culture* magazine in 1989–91 (Plate 2), *Surfer* in 1991–92, and *Ray Gun* magazine in 1992–96—captured the imagination of a youthful audience by taking such an experimental approach into publication design.

Rapid advances in on-screen software also enabled designers to make elements transparent; to stretch, scale, and bend elements; to layer type and images in space; and to combine imagery into complex montages. For example, in a United States postage stamp from 1998, designers Ethel Kessler and Greg Berger digitally montaged John Singer Sargent's portrait of Frederick Law Olmsted with a photograph of New York's Central Park, a site plan, and botanical art to commemorate the landscape architect. Together these images evoke a rich expression of Olmsted's life and work.

The digital revolution in graphic design was followed quickly by public access to the Internet. A whole new area of graphic-design activity mushroomed in the mid-1990s when Internet commerce became a growing sector of the global economy, causing organizations and businesses to scramble to establish Web sites. Designing a Web site involves the layout of screens of information rather than of pages, but approaches to the use of type, images, and colour are similar to those used for print. Web design, however, requires a host of new considerations, including designing for navigation through the site and for using hypertext links to jump to additional information. An example of strong Web design is the Herman Miller for the Home Web site, designed by BBK Studio in 1998. These designers created a strong visual identity, effective navigation, and informational clarity. Attributes that added to the effectiveness of this Web site included a consistent colour palette, an informative use of pictures of products, and a scrolling montage of products.

Web design

Because of the international appeal and reach of the Internet, the graphic-design profession is becoming increasingly global in scope. Moreover, the integration of motion graphics, animation, video feeds, and music into Web-site design has brought about the merging of traditional print and broadcast media. As kinetic media expand from motion pictures and basic television to scores of cable-television channels, video games, and animated Web sites, motion graphics are becoming an increasingly important area of graphic design.

In the 21st century, graphic design is ubiquitous; it is a major component of complex print and electronic information systems. It permeates contemporary society, delivering information, product identification,entertainment, and persuasive messages. The relentless advance of technology has changed dramatically the way graphic designs are created and distributed to a mass audience. However, the fundamental role of the graphic designer—giving expressive form and clarity of content to communicative messages—remains the same.

(P.B.M.)

BIBLIOGRAPHY. Good general sources include STEVEN HELLER and SEYMOUR CHWAST, *Graphic Style from Victorian to Digital* (2000); ALAN LIVINGSTON and ISABELLA LIVINGSTON, *The Thames and Hudson Encyclopaedia of Graphic Design and Designers* (1992); and PHILIP B. MEGGS, *A History of Graphic Design*, 3rd ed. (1998).

Studies on specific aspects of design include NORMA LAFARIE, *The Art & History of Books* (1968, reissued 1995); and JOHN LEWIS, *Anatomy of Printing: The Influences of Art and History on Its Design* (1970). Studies of specific periods of design include JONATHAN J.G. ALEXANDER, *Medieval Illuminators and Their Methods of Work* (1992); ELLEN MAZUR THOMSON, *The Origins of Graphic Design in America, 1870–1920* (1997); and LAUREL HARPER, *Radical Graphics/Graphic Radicals* (1999).

Several important and lavishly illustrated periodicals cover contemporary graphic design and visual communications. *Communication Arts* (8/yr.), published by Coyne & Blanchard; the bimonthly *Graphis*, published by Graphis Inc.; and *Print: America's Graphic Design Magazine*, also bimonthly, published by RC Publications.

(P.B.M.)

Gravitation

Gravitation is a universal force of attraction acting between all matter. It is by far the weakest known force in nature and thus plays no role in determining the internal properties of everyday matter. Due to its long reach and universality, however, gravity shapes the structure and evolution of stars, galaxies, and the entire universe. The trajectories of bodies in the solar system are determined by the laws of gravity, while on Earth all bodies have a weight, or downward force of gravity, proportional to their mass, which the Earth's mass exerts on them. Gravity is measured by the acceleration that it gives to freely falling objects. At the Earth's surface, the acceleration of gravity is about 9.8 metres (32 feet) per second per second. Thus, for every second an object is in free fall, its speed increases by about 9.8 metres per second.

The works of Isaac Newton and Albert Einstein dominate the development of gravitational theory. Newton's classical theory of gravitational force held sway from his *Principia*, published in 1687, until Einstein's work in the early 20th century. Even today, Newton's theory is of sufficient accuracy for all but the most precise applications. Einstein's modern field theory of general relativity predicts only minute quantitative differences from the Newtonian theory except in a few special cases. The major significance of Einstein's theory is its radical conceptual departure from classical theory and its implications for further growth in physical thought.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 131 and 212.

This article is divided into the following sections:

Development of gravitational theory 169
  Early concepts
  Newton's law of gravity
  Potential theory
Acceleration around the Earth, Moon, and planets 171
  Variations in *g*
  Measurements of *g*
  Geophysics
  The Moon and the planets
Gravitational theory and other aspects of physical theory 173
  Field theories of gravitation
  Gravitational fields and general theory of relativity
  The paths of particles and light
  Gravitational radiation
Some astronomical aspects of gravitation 174
Experimental study of gravitation 175
  The inverse square law
  The principle of equivalence
  The constant of gravitation
  The variation of the constant of gravitation with time
  Fundamental character of *G*
Bibliography 177

DEVELOPMENT OF GRAVITATIONAL THEORY

Early concepts. Until Newton's findings, it was not realized that both the movement of celestial bodies and the free fall of objects on Earth are determined by the same force. The classical Greek philosophers, for example, did not consider the celestial bodies to be affected by gravity because the bodies were observed to follow perpetually repeating, non-descending trajectories in the sky. Thus, Aristotle considered that each heavenly body followed a particular "natural" motion, unaffected by external causes or agents. Aristotle also believed that massive earthly objects possess a natural tendency to move toward the Earth's centre. These ideas and two other Aristotelian viewpoints prevailed for centuries: that a body moving at constant speed requires a continuous force acting on it and that force must be applied by contact rather than interaction at a distance. These views impeded the understanding of the principles of motion and precluded the development of ideas about universal gravitation. During the 16th and early 17th centuries, however, several scientific contributions to the problem of earthly and celestial motion set the stage for Newton's later gravitational theory.

Aristotle's views

The 17th-century German astronomer Johannes Kepler accepted the Copernican perspective in which the planets orbit the Sun rather than the Earth. Using the improved measurements of planetary movements made by the Danish astronomer Tycho Brahe during the 16th century, Kepler described the planetary orbits with simple geometric and arithmetic relations. Kepler's three quantitative laws of planetary motion are: (1) the planets describe elliptic orbits, of which the Sun occupies one focus (a focus is one of two points inside an ellipse; rays coming from one of the points bounce off any side of the ellipse and go through the other focus); (2) the line joining a planet to the Sun sweeps out equal areas in equal time; and (3) the square of the period of revolution of a planet is proportional to the cube of its average distance from the Sun. During this same period the Italian astronomer and physicist Galileo made progress in understanding "natural" motion and simple accelerated motion for earthly objects. He realized that bodies that are uninfluenced by forces continue indefinitely to move and that force is necessary to change motion, not to maintain constant motion. In studying how objects fall toward the Earth, Galileo discovered that the motion is one of constant acceleration. He was able to show that the distance a falling body travels from rest in this way varies as the square of the time. As noted above, the acceleration due to gravity at the surface of the Earth is about 9.8 metres per second per second.

Kepler's laws of planetary motion

Newton's law of gravity. Newton discovered the relationship between the motion of the Moon and the motion of any falling body on Earth. His gravitational theory explained Kepler's laws and established the modern quantitative science of gravitation. Newton assumed the presence of an attractive force between all massive bodies, one which does not require bodily contact and acts at a distance. By invoking his law of inertia (bodies not acted upon by a force move at constant speed in a straight line), Newton concluded that a force exerted by the Earth on the Moon is needed to keep it in a circular motion about the Earth rather than moving in a straight line. He realized that this force could be, at long range, the same as the force with which the Earth pulls objects on its surface downward. When Newton discovered that the acceleration of the Moon is $^1/_{3,600}$ smaller than the acceleration at the surface of the Earth, he related the number 3,600 to the radius of the Earth squared. He calculated that the circular orbital motion of radius R and period T requires a constant inward acceleration A equal to the product of $4\pi^2$ and the ratio of the radius to the square of the time:

$$A = \frac{4\pi^2 R}{T^2}. \qquad (1)$$

Newton applied his results to the Moon's orbit, which has a period of 27.3 days and a radius of about 384,000 kilometres (approximately 60 Earth radii). He found the Moon's inward acceleration in its orbit to be 0.0027 metre per second per second, the same as $(^1/_{60})^2$ of the acceleration of a falling object at the surface of the Earth. Because Newton could thus relate the two accelerations to

a common interaction, he deduced that the gravitational force between bodies diminishes as the inverse square of the distance between the bodies. Thus, if the distance between the bodies is doubled, the force on them is reduced to one-quarter. A further result, that the mass of the Earth acts gravitationally on the outside world as if the mass were concentrated at the planet's centre, was needed to obtain his relationship. Newton proved it to be true for all spherically symmetric bodies.

Newton saw that the gravitational force between bodies must depend on the masses of the bodies. Since a body of mass M experiencing a force F accelerates at a rate F/M, a force of gravity proportional to M would be consistent with Galileo's observation that all bodies accelerate under gravity toward the Earth at the same rate. In Newton's equation, F_{12} is the magnitude of the gravitational force acting between masses M_1 and M_2 separated by distance r_{12}. The force equals the product of these masses and of G, a universal constant, divided by the square of the distance.

The universal constant G

$$F_{12} = \frac{GM_1M_2}{r_{12}^2} \tag{2}$$

The constant G is a quantity with the physical dimensions $(\text{length})^3/(\text{mass})(\text{time})^2$; its numerical value depends on the physical units of length, mass, and time used. (G is discussed more fully in subsequent sections.)

The force acts in the direction of the line joining the two bodies and so is represented naturally as a vector, $\boldsymbol{F}$. If $\boldsymbol{r}$ is the vector separation of the bodies, then

$$\boldsymbol{F} = \frac{GM_1M_2\boldsymbol{r}}{r^3}. \tag{3}$$

In this expression, the factor $\boldsymbol{r}/r^3$ acts in the direction of $\boldsymbol{r}$ and is numerically equal to $1/r^2$.

The attractive force of a number of bodies of masses M_i on a body of mass M is

$$F = \frac{GM\Sigma_i M_i r_i}{r_i^3}, \tag{4}$$

where Σ_i means that the vector forces due to all the attracting bodies must be added together. This is Newton's gravitational law essentially in its original form. A simpler expression, equation (5), gives the surface acceleration on Earth. Setting a mass equal to the Earth's mass M_E and the distance equal to the Earth's radius r_E, the downward acceleration of a body at the surface g is equal to the product of the universal gravitational constant and the mass of the Earth divided by the square of the radius:

$$g = \frac{GM_E}{r_E^2}. \tag{5}$$

Weight and mass

The weight W of the body can be measured by the equal and opposite force necessary to prevent the downward acceleration; this is M_i. The same body placed on the surface of the Moon has the same mass, but, as the Moon has a mass of about $^1/_{81}$ times that of the Earth and a radius of just 0.27 that of the Earth, the body on the lunar surface has a weight of only $^1/_6$ its Earth weight, as the U.S. Apollo astronauts demonstrated. Passengers and instruments in orbiting satellites, where no force prevents the free fall of the satellites in the gravitational field, experience weightless conditions even though their masses remain the same as on Earth.

Equations (1) and (2) above can be used to derive Kepler's third law for the case of circular planetary orbits. By using the expression for the acceleration A in equation (1) for the force of gravity for the planet GM_PM_S/R^2 divided by the planet's mass M_P, the following equation, in which M_S is the mass of the Sun, is obtained:

$$\frac{GM_S}{R^2} = \frac{4\pi^2 R}{T^2}$$

or

$$R^3 = \left(\frac{GM_S}{4\pi^2}\right)T^2. \tag{6}$$

Newton was able to show that all three of Kepler's observationally derived laws follow mathematically from the assumption of his own laws of motion and gravity. In all observations of the motion of a celestial body, only the product of G and the mass can be found. Newton first estimated the magnitude of G by assuming the Earth's average mass density to be about 5.5 times that of water (somewhat greater than the Earth's surface rock density) and by calculating the Earth's mass from this. Then, taking M_E and r_E as the Earth's mass and radius, respectively, the value of G was

$$G = \frac{gr_E^2}{M_E}, \tag{7}$$

which numerically comes close to the accepted value of 6.6726×10^{-11} $\text{m}^3\text{s}^{-2}/\text{kg}^{-1}$, first directly measured by a Cavendish balance experiment (see below).

Comparing equation (5) above for the Earth's surface acceleration g with the R^3/T^2 ratio for the planets, a formula for the ratio of the Sun's mass, M_S, to Earth's mass, M_E, was obtained in terms of known quantities, R_E being the radius of the Earth's orbit:

$$\frac{M_S}{M_E} = \frac{4\pi^2 R_E^3}{gT_E^2 r_E^2} \cong 325{,}000. \tag{8}$$

Using observations of the motion of the moons of Jupiter discovered by Galileo, Newton determined that Jupiter was 318 times more massive than the Earth but only $^1/_4$ as dense, having a radius 11 times larger than the Earth's.

Interaction among celestial bodies

When two celestial bodies of comparable mass interact gravitationally, both orbit about a fixed point (the centre of mass of the two bodies). This point lies between the bodies on the line joining them at a position such that the distances to each body multiplied by each body's mass are equal. Thus, the Earth and the Moon move in an orbit about their common centre of mass. This motion of the Earth has two observable consequences. First, the direction of the Sun as seen from the Earth relative to the very distant stars varies each month by about 12 arc seconds in addition to the Sun's annual motion. Second, the line-of-sight velocity from the Earth to a freely moving spacecraft varies each month by 2.04 metres per second according to very accurate data obtained from radio tracking. From these results the Moon is found to have a mass $^1/_{81}$ times that of the Earth. With slight modifications Kepler's laws remain valid for systems of two comparable masses; the foci of the elliptical orbits are the two-body centre-of-mass positions, and, putting $M_1 + M_2$ instead of M_S in the expression of Kepler's third law, equation (6) above, the third law reads:

$$R^3 = \frac{G(M_1 + M_2)}{4\pi^2}T^2. \tag{9}$$

This agrees with equation (6) when one body is so small that its mass can be neglected. The rescaled formula can be used to determine the separate masses of binary stars (pairs of stars orbiting around each other) that are a known distance from the solar system. Equation (9) determines the sum of the masses; and, if R_1 and R_2 are the distances of the individual stars from the centre of mass, the ratio of the distances must balance the inverse ratio of the masses, and the sum of the distances is the total distance R. In symbols,

$$\frac{R_1}{R_2} = \frac{M_2}{M_1};\ R_1 + R_2 = R. \tag{10}$$

These relations are sufficient to determine the individual masses. Observations of the orbital motion of double stars, of the dynamic motion of stars collectively moving within their galaxies, and of the motion of the galaxies themselves verify that Newton's law of gravity is valid to a high degree of accuracy throughout the visible universe.

Ocean tides, phenomena that mystified thinkers for centuries, were also shown by Newton to be a consequence of the universal law of gravitation, although the details of the complicated phenomena were not understood until comparatively recently. They are caused specifically by the gravitational pull of the Moon and, to a lesser extent, of the Sun (see OCEANS: *Tide-generating forces*).

It was already known in Newton's day that the Moon does not move in a simple Keplerian orbit. Later, more accurate observations of the planets also showed discrep-

ancies from Kepler's laws. The motion of the Moon is particularly complex; however, apart from a long-term acceleration due to tides on the Earth, it can be accounted for by the gravitational attraction of the Sun, the Earth, and the other planets. In addition, the gravitational attraction of the planets for each other explains almost all the features of their motions. The exceptions are nonetheless important. Uranus, the seventh planet from the Sun, was observed to undergo variations in its motion that could not be explained by perturbations from Saturn, Jupiter, and the other planets. Two 19th-century astronomers, John Couch Adams of Britain and Urbain-Jean-Joseph Le Verrier of France, independently assumed the presence of an unseen eighth planet that could produce the observed discrepancies. They calculated its position within a degree of which the planet Neptune was discovered in 1846. Measurements of the motion of the innermost planet, Mercury, over an extended period led astronomers to conclude that the major axis of this planet's elliptical orbit precesses in space at a rate 43 arc seconds per century faster than could be accounted for from perturbations of the other planets. In this case, however, no other bodies could be found that could produce this discrepancy, and very slight modification of Newton's law of gravitation seemed to be needed. Einstein's theory of relativity precisely predicts this observed behaviour of Mercury's orbit (see RELATIVITY).

The discovery of Neptune

Potential theory. For irregular, nonspherical mass distributions in three dimensions, Newton's original vector equation (4, above) is inefficient, though theoretically it could be used for finding the resulting gravitational field. The main progress in classical gravitational theory after Newton was the introduction of potential theory, which is the mathematical representation of gravitational fields. It allows practical as well as theoretical investigation of the gravitational variations in space and of the anomalies due to the irregularities and shape deformations of the Earth.

Potential theory led to the following elegant formulation: the gravitational acceleration, g, a function of position $\boldsymbol{R}$, $\boldsymbol{g}(\boldsymbol{R})$, at any point in space is given from a function, Φ, called the gravitational potential, by means of a generalization of the operation of differentiation:

$$\boldsymbol{g}(\boldsymbol{R}) = \frac{\partial \Phi}{\partial x}\boldsymbol{i} + \frac{\partial \Phi}{\partial y}\boldsymbol{j} + \frac{\partial \Phi}{\partial z}\boldsymbol{k}, \tag{11}$$

in which $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ stand for unit basis vectors in a three-dimensional Cartesian coordinate system. The potential and therefore g are determined by an equation discovered by the French mathematician Siméon-Denis Poisson:

Poisson's equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\Phi(\boldsymbol{R}) = -4\pi G\rho(\boldsymbol{R}). \tag{12}$$

The significance of this approach is that Poisson's equation can be solved under rather general conditions, which is not the case with Newton's equation. When mass density ρ is nonzero, the solution is expressed as a definite integral:

$$\Phi(\boldsymbol{R}) = G\int \frac{\rho(\boldsymbol{R}')d\boldsymbol{R}'}{|\boldsymbol{R} - \boldsymbol{R}'|}. \tag{13}$$

When $\rho = 0$ (in particular, outside the Earth), Poisson's equation reduces to the simpler equation of Laplace.

The appropriate coordinates for the region outside the nearly spherical Earth are spherical polar coordinates: $\boldsymbol{R}$, the distance from the centre of the Earth; θ, the colatitude measured from the North Pole; and the longitude measured from Greenwich. The solutions are series of powers of trigonometric functions of colatitude and longitude, known as spherical harmonics; the first terms are:

$$\Phi(R) = -\frac{GM_E}{R}\left[1 - J_2\left(\frac{R_E}{R}\right)^2 \frac{3\cos^2\theta - 1}{2}\right.$$
$$\left. - J_3\left(\frac{R_E}{R}\right)^3 \frac{5\cos^3\theta - 3\cos\theta}{2} + \ldots\right], \tag{14}$$

The constants J_2, J_3, and so forth are determined by the detailed mass distribution of the Earth; and, since Newton showed that for a spherical body all the J_n are zero, they must measure the deformation of the Earth from a spherical shape. J_2 measures the magnitude of the Earth's rotational equatorial bulge, J_3 measures a slight pear-shaped deformation of the Earth, and so on. From observations of perturbations on satellite orbits, the parameters J_2 and J_3 have been found to be $1{,}082.7 \times 10^{-6}$ and -2.4×10^{-6}, respectively.

Much the largest variation in the potential around the Earth is due to the equatorial bulge, and there is a corresponding increase in the value of gravity on the surface of the Earth from the Equator to the poles. Newton was the first to give a theory of the equatorial bulge; among the data he used to estimate its size were some pendulum measurements made at Saint Helena by the English astronomer Edmond Halley (see below). Many more harmonic components have been estimated from observations of the orbits of artificial satellites close to Earth; the sum of the most important of them can be shown by contours on a map of Earth's surface at sea level, where the potential is a constant (see Figure 1).

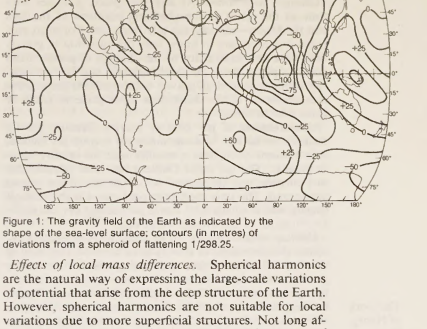


Figure 1: The gravity field of the Earth as indicated by the shape of the sea-level surface; contours (in metres) of deviations from a spheroid of flattening 1/298.25

Effects of local mass differences. Spherical harmonics are the natural way of expressing the large-scale variations of potential that arise from the deep structure of the Earth. However, spherical harmonics are not suitable for local variations due to more superficial structures. Not long after Newton's time, it was found that the gravity on top of large mountains is less than expected on the basis of their visible mass. The idea of isostasy was developed, according to which the unexpectedly low acceleration of gravity on a mountain is caused by low-density rock 30 to 100 kilometres underground, which buoys up the mountain. Correspondingly, the unexpectedly high force of gravity on ocean surfaces is explained by dense rock 30 to 100 kilometres beneath the ocean bottom.

Concept of isostasy

Portable gravimeters, which can detect variations of one part in 10^8 in the gravitational force, are in wide use today for mineral and oil prospecting. Unusual underground deposits reveal their presence by producing local gravitational variations (see below).

Weighing the Earth. Because the mass of the Earth can be calculated from g and the planet's radius if G is known, the English physicist-chemist Henry Cavendish and other early experimenters termed their work "weighing the Earth." The mass of the Earth is about 5.98×10^{24} kilograms, while the mean densities of the Earth, Sun, and Moon are, respectively, 5.52, 1.43, and 3.3 times that of water. (K.L.N./A.H.C.)

ACCELERATION AROUND THE EARTH, MOON, AND PLANETS

The value of the attraction of gravity or of the potential is determined by the distribution of matter within the Earth or other celestial body. In turn, the distribution of matter determines the shape of the surface on which the potential is constant (Figure 1). Measurements of gravity and the potential are thus essential both to geodesy, which is the study of the shape of the Earth, and to geophysics, the

study of its internal structure. For geodesy and global geophysics, it is best to measure the potential from the orbits of artificial satellites. Surface measurements of gravity are best for local geophysics, which deals with the structure of mountains and oceans and the search for minerals.

Variations in g. *Changes due to location.* The acceleration g varies by about $\frac{1}{2}$ of 1 percent with position on the Earth's surface, from about 9.78 metres per second per second at the Equator to approximately 9.83 metres per second per second at the poles. This variation stems chiefly from the rotation of the Earth, as part of the Earth's pull is balanced by keeping objects rotating with the Earth instead of flying tangentially off into space (as mud does off a spinning wheel). This effect is also responsible for the bulge of the Earth at the Equator and the slight flattening at the poles. The distance to the centre of the Earth, therefore, increases with the bulge from the poles to the Equator, and consequently the value of g, which for a spherical Earth of radius r_E and mass M_E is given simply by GM_E/r_E^2, is less toward the Equator.

In addition to this broad-scale variation, local variations of a few parts in 10^6 or smaller are caused by variations in the density of the Earth's crust as well as height above sea level.

Changes with time. At any one place, g varies with time as the result of the changing positions of the Sun and the Moon—namely, the tidal variation. For most purposes it is necessary to know only the variation of gravity with time at a fixed place or the changes of gravity from place to place; then, the tidal variation can be removed. Accordingly, almost all gravity measurements are relative measurements of the differences from place to place or from time to time.

Measurements of g. *Unit of gravity.* Because gravity changes are far less than one metre per second per second, it is convenient to have a smaller unit for relative measurements. The gal (after Galileo) has been adopted for this purpose; one gal equals $\frac{1}{100}$ of one metre per second per second. The unit most commonly used is the milligal (mgal), which equals 10^{-5} metre per second per second—*i.e.*, $\frac{1}{1,000,000}$ of the average value of g.

Absolute measurements. Two basic ways of making absolute measurements of gravity have been devised: timing the free-fall of an object and timing the motion under gravity of a body constrained in some way, almost always by a pendulum. In 1817 the English physicist Henry Kater, building on the work of the German astronomer Friedrich Wilhelm Bessel, was the first to use a reversible pendulum to make absolute measurements of g. If the periods of swing of a rigid pendulum about two alternative points of support are the same, then the separation of those two points is equal to the length of the equivalent simple pendulum of the same period. By careful construction, Kater was able to measure the separation very accurately. The so-called reversible pendulum was used for absolute measurements of gravity from Kater's day until the 1950s. Since that time, electronic instruments have enabled investigators to measure with high precision the half-second time of free-fall of a body (from rest) through one metre. It is also possible to make extremely accurate measurements using light interference. Consequently, direct measurements of free-fall have replaced the pendulum for absolute measurements of gravity. Nowadays lasers serve as light sources for interferometers; the falling object, in effect, becomes a retroreflector that returns a beam of light back upon itself. Transportable versions of such apparatus have been used in different locations to establish a basis for measuring differences of gravity over the entire Earth. The accuracy attainable in these measurements is about one part in 10^8.

The work of Henry Kater

Relative measurements. From the time of Newton, measurements of differences of gravity (strictly, the ratios of values of gravity) were made by timing the same pendulum at different places. During the 1930s, however, static gravimeters replaced pendulums for local measurements over small ranges of gravity. Today, free-fall measurements have rendered the pendulum obsolete for all purposes.

Spring gravimeters balance the force of gravity mg on a mass m in the gravity field to be measured, against the elastic force of the spring, using either electronic or mechanical means to achieve high sensitivity. Vibrating string gravimeters in which the string's vibration frequency is determined by g also have been developed. A device of this type was employed by the Apollo 17 astronauts on the Moon to conduct a gravity survey of their landing site. Another relatively recent development is the superconducting gravimeter, an instrument in which the position of a magnetically levitated superconducting sphere is sensed to provide a measure of g.

Spring gravity meters

Modern gravimeters may have sensitivities better than 0.005 milligal, the standard deviation of observations in exploration surveys being on the order of 0.01–0.02 milligal.

Differences in gravity measured with gravimeters are obtained in quite arbitrary units—divisions on a graduated dial, for example. The relation between these units and milligals can only be determined by reading the instrument at a number of points where g is known as a result of absolute or relative pendulum measurements. Further, because an instrument will not have a completely linear response, known points must cover the entire range of gravity over which the gravimeter is to be used.

Gravimetric surveys. Recently, by combining all available absolute and relative measurements, it has been possible to obtain the most probable gravity values at a large number of sites to a high degree of accuracy. The culmination of gravimetric work begun in the 1960s has been a worldwide gravity reference system having an accuracy of at least one part in 10^7 (0.1 milligal or better).

Since g is an acceleration, the problem of its measurement from a vehicle that is moving and therefore unavoidably accelerating relative to the Earth raises a number of fundamental problems. Pendulum, vibrating-string, and spring-gravimeter observations have been made from submarines; using gyrostabilized platforms, relative gravity measurements with accuracies approaching a few milligals have been and are being made from surface ships. Experimental measurements with various gravity sensors on fixed-wing aircraft as well as on helicopters have been carried out.

Geophysics. The value of gravity measured at the terrestrial surface is the result of (1) the gravitational attraction of the Earth as a whole, (2) centrifugal force caused by the Earth's rotation, (3) elevation, (4) unbalanced attractions caused by surface topography, (5) tidal variations, and (6) unbalanced attractions caused by irregularities in underground density distributions. Most geophysical surveys are aimed at separating out the last of these in order to interpret the geologic structure. It is therefore necessary to make proper allowance for the other factors.

The first two factors imply a variation of gravity with latitude that can be calculated for an assumed shape for the Earth. The third factor, which is the decrease in gravity with elevation, due to increased distance from the centre of the Earth, amounts to -0.3086 milligal per metre. This value, however, assumes that material of zero density occupies the whole space between the point of observation and sea level, and it is therefore termed the free-air correction factor. In practice, the mass of rock material that occupies part or all of this space must be considered. In an area where the topography is reasonably flat, this is usually calculated by assuming the presence of an infinite slab of thickness equal to the height of the station, h, and having an appropriate density σ: its value is $+0.04185\ \sigma h$ milligal per metre. This is commonly called the Bouguer correction factor.

The free-air and Bouguer corrections

Terrain or topographic corrections also can be applied to allow for the attractions due to surface relief if the densities of surface rocks are known. Tidal effects (the amplitudes are less than 0.3 milligal) can be calculated and allowed for.

(Ja.F./A.H.C.)

The Moon and the planets. It was noted above that the large-scale features of the Earth's potential are obtained from observations of Earth-orbiting satellites and may be shown as a map of the sea-level surface (Figure 1). Although the Apollo astronauts used a gravimeter at their lunar landing site, most scientific knowledge about the Moon's gravitational attraction has been derived from observations of its effects on the accelerations of spacecraft. Information about the gravitational force of various plan-

ets has been obtained in this way as well. Radio tracking makes it possible to estimate the accelerations of spacecraft very accurately, and the results can be expressed either as terms in a series of spherical harmonics or as the variation of gravity over the surface. As in the case of the Earth, spherical harmonics are more effective for studying gross structure, while the variation of gravity is more useful for local features. Because spacecraft must descend close to the surface to detect local gravity variations, only data for the Moon and Mars have been obtained so far.

The Moon's polar flattening is much less than that of the Earth, while its equator is far more elliptical. There are also large, more local irregularities from visible and concealed structures. Mars also exhibits some large local variations, while the equatorial bulges of Mercury and Venus are very slight. By contrast, the major planets, all of which rotate quite fast, have large equatorial bulges, and their gravity is dominated by a large increase from equator to pole. These results are crucial to understanding the internal properties of the planets.

GRAVITATIONAL THEORY AND OTHER ASPECTS OF PHYSICAL THEORY

The Newtonian theory of gravity is based on an assumed force acting between all pairs of bodies—*i.e.*, an action at a distance. When a mass moves, the force acting on other masses has been considered to adjust instantaneously to the new location of the displaced mass. Special relativity theory states that no physical signal travels faster than the speed of light and that all signals travel at this speed through empty space. This theory, with the field theory of electrical and magnetic phenomena, has met such empirical success that most modern gravitational theories are constructed as field theories consistent with the principles of special relativity. In a field theory the gravitational force between bodies is formed by a two-step process: (1) One body produces a gravitational field that permeates all surrounding space but has weaker strength farther from its source. A second body in that space is then acted upon by this field and experiences a force. (2) The Newtonian force of reaction is then viewed as the response of the first body to the gravitational field produced by the second body, there being at all points in space a superposition of gravitational fields due to all the bodies in it.

Special relativity and electromagnetic field theory

Field theories of gravitation. The possibility that gravitation might be linked with the other forces of nature in a unified theory of forces greatly increased interest in gravitational field theories during the 1970s and '80s. The first such unified field theory, and so far the only successful one, is that of physicists Abdus Salam of Pakistan and Steven Weinberg and Sheldon L. Glashow of the United States, who proposed that the electromagnetic forces and the weak force responsible for beta decay are different manifestations of one and the same basic interaction. Physicists now are actively seeking other possible unified combinations. Because the gravitational force is exceedingly weak compared with all others and because it seems to be independent of all physical properties except mass, the unification of gravitation with the other forces will be the most difficult to achieve. This challenge has provided a tremendous impetus to experimental investigations to determine whether there may be some failure of the apparent independence.

The prime example of a field theory is Einstein's general relativity, according to which the acceleration due to gravity is a purely geometric consequence of the properties of space-time in the neighbourhood of attracting masses. (As will be seen below, general relativity makes certain specific predictions that are borne out well by observation.) In a whole class of more general theories, these and other effects not predicted by simple Newtonian theory are characterized by free parameters; such formulations are called parameterized post-Newtonian (PPN) theories. There is now considerable experimental and observational evidence for limits to the parameters. So far, no deviation from general relativity has been demonstrated convincingly.

Field theories of gravity predict specific corrections to the Newtonian force law, the corrections taking two basic forms: (1) When matter is in motion, additional gravitational fields (analogous to the magnetic fields produced by moving electric charges) are produced; also, moving bodies interact with gravitational fields in a motion-dependent way. (2) Unlike electromagnetic field theory, in which two or more electric or magnetic fields superimpose by simple addition to give the total fields, in gravitational field theory nonlinear fields proportional to the second and higher power of the source masses are generated, and gravitational fields proportional to the product of different masses are created. Gravitational fields themselves become sources for additional gravitational fields. Examples of some of these effects are shown below. The acceleration, A, of a moving particle of negligible mass that interacts with a mass, M, which is at rest, is given in the following formula, derived from Einstein's gravitational theory. The expression for A now has, as well as the Newtonian expression from equation (1), further terms in higher powers of GM/R^2—that is, in G^2M^2/R^4. As elsewhere, V is the particle's velocity vector, A is its acceleration vector, R is the vector from the mass M, and c is the speed of light. When written out, the sum is

Corrections of Newtonian force law

$$A = -\frac{GMR}{R^3} + 2\frac{G^2M^2R}{c^2R^2} - \frac{3}{2}\frac{GMR}{R^3}\left(\frac{V^2}{c^2}\right) - \frac{V\cdot AV}{c^2} - \frac{1}{2}\frac{V^2}{c^2}A + \ldots \tag{15}$$

This expression gives only the first post-Newtonian corrections; terms of higher power in $1/c$ are neglected. For planetary motion in the solar system the $1/c^2$ terms are smaller than Newton's acceleration term by at least a factor of 10^{-8}, but some of the consequences of these correction terms are measurable and important tests of Einstein's theory. It should be pointed out that prediction of new observable gravitational effects requires particular care; Einstein's pioneer work in gravity has shown that gravitational fields affect the basic measuring instruments of experimental physics—clocks, rulers, light rays—with which any experimental result in physics is established. Some of these effects are listed below:

1. The rate at which clocks run is reduced by proximity of massive bodies; *i.e.*, clocks near the Sun will run slowly compared with identical clocks farther away from it.

2. In the presence of gravitational fields the spatial structure of physical objects is no longer describable precisely by Euclidean geometry; for example, in the arrangement of three rigid rulers to form a triangle, the sum of the subtended angles will not equal 180°. A more general type of geometry, Riemannian geometry, seems required to describe the spatial structure of matter in the presence of gravitational fields.

3. Light rays do not travel in straight lines, the rays being deflected by gravitational fields. To distant observers the light-propagation speed is observed to be reduced near massive bodies.

Gravitational fields and general theory of relativity. In Einstein's general theory of relativity the physical consequences of gravitational fields are stated in the following way. Space-time is a four-dimensional non-Euclidean continuum, the curvature of space-time's Riemannian geometry being produced by or related to the world's matter distribution. Particles and light rays travel along the geodesics (shortest paths) of this four-dimensional geometric world.

There are two principal consequences of the geometric view of gravitation: (1) the accelerations of bodies depend only on their masses and not on their chemical or nuclear constitution, and (2) the path of a body or of light in the neighbourhood of a massive body (the Sun, for example) is slightly different from that predicted by Newton's theory. The first is the weak principle of equivalence. Newton himself performed experiments with pendulums that demonstrated the principle to better than one part in 1,000 for a variety of materials, and, at the beginning of the 20th century, the Hungarian physicist Roland, Baron von Eötvös, showed that different materials accelerate in the Earth's field at the same rate to within one part in 10^9. More recent experiments have shown the equality of

Weak principle of equivalence

accelerations in the field of the Sun to within one part in 10^{11}. Newtonian theory is in accord with these results because of the postulate that gravitational force is proportional to a body's mass.

Inertial mass is a mass parameter giving the inertial resistance to acceleration of the body when responding to all types of force. Gravitational mass is determined by the strength of the gravitational force experienced by the body when in the gravitational field g. The Eötvös experiments, therefore, show that the ratio of gravitational and inertial mass is the same for different substances.

Einstein's special theory of relativity views inertial mass as a manifestation of all the forms of energy in a body according to his fundamental relationship $E = mc^2$, E being the total energy content of a body, m the inertial mass of the body, and c the speed of light. Dealing with gravitation, then, as a field phenomenon, the weak principle of equivalence indicates that all forms of nongravitational energy must identically couple to or interact with the gravitational field, because the various materials in nature possess different fractional amounts of nuclear, electrical, magnetic, and kinetic energies, yet they accelerate at identical rates.

In the general theory of relativity the gravitational field also interacts with gravitational energy in the same manner as with other forms of energy, an example of that theory's universality not possessed by most other theories of gravitation.

The Sun has an appreciable fraction of internal gravitational energy, and the repetitions of the Eötvös experiments during the 1970s with the Sun instead of the Earth as the attracting mass revealed that bodies accelerate at identical rates in the Sun's field as well as in that of the Earth. Extremely accurate laser measurements of the distance of the Moon from the Earth have made possible a further test of the weak principle of equivalence. The chemical constitutions of the Earth and the Moon are not the same, and so, if the principle did not hold, they might accelerate at different rates under the Sun's attraction. No such effect has been detected.

Newton's third dynamic law states that every force implies an equal and opposite reaction force. Modern field theories of force contain this principle by requiring every entity that is acted upon by a field to be also a source of the field. An experiment by the American physicist Lloyd Kreuzer established to a one-part-in-20,000 accuracy that different materials produce gravitational fields with a strength the same as that of gravitational fields acting upon them. In this experiment a sphere of solid material was moved through a liquid of identical weight density. The absence of a gravitational effect on a nearby Cavendish balance instrument during the sphere's motion is interpreted as showing that the two materials had equal potency in producing a local gravitational-field anomaly.

The gravitational redshift

Other experiments have brought confirmation of Einstein's predictions to an accuracy of a few percent. Using the Mössbauer effect to monitor the nuclear reabsorption of resonant gamma radiation, a shift of wavelength of the radiation that traveled vertically tens of metres in the Earth's gravitational field was measured, and the slowing of clocks (in this case the nuclear vibrations are clocks) as predicted by Einstein was confirmed to 1 percent precision. If ν and $\Delta\nu$ are clock frequency and change of frequency, respectively, h is the height difference between clocks in the gravitational field g. Then

$$\frac{\Delta\nu}{\nu} = -\frac{gh}{c^2}. \tag{16}$$

For a height of 10 metres, there is only a one-part-in-10^{15} change in clock rates.

The paths of particles and light. The idea that light should be deflected by passing close to a massive body had been suggested by the British astronomer and geologist John Michell in the 18th century. However, Einstein's general relativity theory predicted twice as much deflection as Newtonian physics. Quick confirmation of Einstein's result came from measuring the direction of a star close to the Sun during an expedition led by the British astronomer Sir Arthur Stanley Eddington to observe the solar eclipse of 1919. Optical determinations of the change of direction of a star are subject to many systematic errors, and far better confirmation of Einstein's general relativity theory has been obtained from measurements of a closely related effect—namely, the increase of the time taken by electromagnetic radiation along a path close to a massive body. In fact, both effects may be understood as due to a decrease in the speed of light near a massive object. Timing the round-trip travel time for radar pulses between the Earth and other inner planets or artificial satellites passing behind the Sun, experiments have confirmed to about 4 percent the prediction of an additional time delay, Δt. This additional time delay is given by the following formula, in which M_S is the Sun's mass, R_1 and R_2 are the distances from the Sun to Earth and to the other reflecting body, and D is the distance of closest approach of the radar pulses to the Sun (ln stands for natural logarithm):

$$\Delta t = \frac{4GM_S}{c^3} \ln \frac{4R_1R_2}{D^2}. \tag{17}$$

The additional precession of the orbit of Mercury of 43 arc seconds per century (see above) was known before the development of the theory of general relativity. With radar measurements of the distances to the planets, similar anomalous precessions have been estimated for Venus and the Earth and have been found to agree with general relativity.

Gravitational radiation. An important consequence of field theories of gravitation is that the gravitational field itself can oscillate in the same way that an electromagnetic field can. Thus, if the masses that are the source of a field change with time, they may be expected to radiate energy through waves in the field. The radiation would be very weak, and, though some extremely sensitive instruments have been built in an effort to detect gravitational radiation, none has been observed with certainty so far. The principal sources are expected to be double stars in their regular motions and massive stars collapsing as supernovas. Even though no gravitational radiation has yet been detected on Earth, there are strong grounds for believing that it exists. One particular double star system has a pulsar as one of its components, and, from measurements of the Doppler shifts of the pulsar frequency, precise estimates of the period of the orbit show that the period is changing, corresponding to a decrease in the energy of the orbital motion. Gravitational radiation is the only known means by which this could happen.

Possible sources of gravitational radiation

(K.L.N./A.H.C.)

SOME ASTRONOMICAL ASPECTS OF GRAVITATION

As stated above, studies of gravity enable researchers to determine the masses and densities of celestial bodies and thereby make it possible to investigate the physical constitutions of stars and planets. Because gravitation is a very weak force, however, its distinctive effects appear only when masses are extremely large. The idea that light might be attracted gravitationally had been suggested by the aforementioned John Michell and examined by the French mathematician and astronomer Pierre-Simon Laplace. Predictions by classical physics and general relativity that light passing close to the Sun might be deflected are described above. There are two further consequences for astronomy. Light from a distant object may pass close to objects other than the Sun and be deflected by them. In particular, they may be deflected by a massive galaxy. If some object is behind a massive galaxy, as seen from Earth, deflected light may reach the Earth by more than one path. Operating like a lens that focuses light along different paths, the gravity of the galaxy may make the object appear multiple; examples of such apparently double objects have been found.

Both Michell and Laplace pointed out that the attraction of a very dense object upon light might be so great that the light could never escape from the object, rendering it invisible. Such a phenomenon is a black hole. The relativistic theory of black holes has been thoroughly developed in recent years, and astronomers have conducted an intense search for them. One possible class of black holes comprises stars that have used up all their nuclear

Black holes

energy so that they are no longer held up by radiation pressure and have collapsed to neutron stars of extremely high density. Black holes also are thought to be at the centres of the larger galaxies and to be the sources of the energy found there.

Black holes from which no radiation is able to escape cannot be seen by their own light, but there may be observable secondary effects. If a black hole were one component of a double star, the orbital motion of the pair and the mass of the invisible member might be derived from the oscillatory motion of a visible companion. Because black holes attract matter, any gas in the vicinity of an object of this kind would fall into it and acquire, before vanishing into the hole, a high velocity and consequently a high temperature. The gas may become hot enough to produce X rays and gamma rays from around the hole. While there is still no definite proof, such a mechanism may be the origin of at least some powerful X-ray and radio astronomical sources, including those at the centres of galaxies and quasars (see COSMOS: *Black-hole model for active galactic nuclei*). Only astronomical objects can be expected to be sources of detectable gravitational radiation. As already mentioned, gravitational radiation is probably responsible for secular changes in the orbits of some double stars, and so in the very long term it may have an effect on the stability of celestial objects. If and when gravitational radiation is detected, new astronomical phenomena will no doubt be discovered.

EXPERIMENTAL STUDY OF GRAVITATION

The essence of Newton's theory of gravitation is that the force between two bodies is proportional to the product of their masses and the inverse square of their separation and that the force depends on nothing else. With a small geometric modification, the same is true in general relativity. Newton himself tested his assumptions by experiment and observation. He made pendulum experiments to confirm the principle of equivalence and checked the inverse square law as applied to the periods and diameters of the orbits of the satellites of Jupiter and Saturn. During the latter part of the 19th century many experiments showed the force of gravity to be independent of temperature, electromagnetic fields, shielding by other matter, orientation of crystal axes, and other factors. The revival of such experiments during the 1970s was the result of theoretical attempts to relate gravitation to other forces of nature by showing that general relativity incompletely describes gravity. New experiments on the equivalence principle were performed, and experimental tests of the inverse square law were made in the laboratory. There also has been a continuing interest in the determination of the constant of gravitation, G, although it must be pointed out that G occupies a rather anomalous position among the other constants of physics. In the first place, the mass, M, of any celestial object cannot be determined independently of the gravitational attraction that it exerts. Thus, the combination GM, not the separate value of M, is the only meaningful property of a star, planet, or galaxy. Second, according to general relativity and the principle of equivalence, G does not depend on material properties but is in a sense a geometric factor. Hence, the determination of the constant of gravitation does not seem as essential as the measurement of quantities like the electronic charge or Planck's constant. It is also much less well determined experimentally than any of the other constants of physics.

The constant of gravitation as a geometric factor

Experiments on gravitation are in fact very difficult, as a comparison of experiments on the inverse square law of electrostatics with those on gravitation will show. The electrostatic law has been established to within one part in 10^{16} by using the fact that the field inside a closed conductor is zero when the inverse square law is in effect and then detecting any residual fields with very sensitive electronic devices. Gravitational forces have to be detected by mechanical means, most often the torsion balance, and, though the sensitivities of mechanical devices have been greatly improved, they are still far less sensitive than electronic equipment. Mechanical arrangements also preclude the use of a complete gravitational enclosure. Last, extraneous disturbances are relatively large because gravitational forces are very small (something that Newton first pointed out). Thus, the inverse square law is established over laboratory distances to no better than one part in 10^4.

The inverse square law. The current interest in the inverse square law arises from two suggestions. First, the gravitational field itself might have a mass, in which case the constant of gravitation would change in an exponential manner from one value for small distances to a different one for large distances over a characteristic distance related to the mass of the field. Second, the observed field might be the superposition of two or more fields of different origin and different strengths, one of which might depend on chemical or nuclear constitution. Deviations from the inverse square law have been sought in three ways: (1) the law has been checked in the laboratory over distances up to about one metre, (2) the effective value of G for distances between 100 metres and one kilometre has been estimated from geophysical studies, and (3) there have been careful comparisons of the value of the attraction of the Earth as measured on the surface and as experienced by artificial satellites.

Search for deviations from the inverse square law

Early in the 1970s an experiment by the American physicist Daniel R. Long seemed to show deviation from the inverse square law at a range of about 0.1 metre. Long compared the maximum attractions of two rings upon a test mass hung from the arm of a torsion balance. The maximum attraction of a ring occurs at a particular point on the axis and is determined by the mass and dimensions of the ring. If the ring is moved until the force on the test mass is greatest, the distance between the test mass and the ring is not needed. Two later experiments over the same range showed no deviation from the inverse square law. In one, conducted by the American physicist Riley Newman and his colleagues, a test mass hung on a torsion balance was moved around in a long hollow cylinder. The cylinder approximates a complete gravitational enclosure and, allowing for a small correction because it is open at the ends, the force on the test mass should not depend on its location within the cylinder. No deviation from the inverse square law was found. In the other experiment, performed in Cambridge, Eng., by Y.T. Chen and associates, the attractions of two solid cylinders of different mass were balanced against a cylinder with a third mass so that only the separations of the cylinders had to be known; it was not necessary to know the distances of any from a test mass. Again no deviation of more than one part in 10^4 from the inverse square law was found. Other somewhat less sensitive experiments at ranges up to one metre or so also have failed to establish any greater deviation.

The geophysical tests go back to a method for the determination of the constant of gravitation that had been used in the 19th century, especially by the British astronomer Sir George Airy. Suppose the value of gravity, g, is measured at the top and bottom of a horizontal slab of rock of thickness t and density d. The values for the top and bottom will be different for two reasons. First, the top of the slab is t farther from the centre of the Earth, and so the measured value of gravity will be less by $2(t/R)g$, where R is the radius of the Earth. Second, the slab itself attracts objects above and below it toward its centre; the difference between the downward and upward attractions of the slab is $4\pi Gtd$. Thus, a value of G may be estimated. Frank D. Stacey and his colleagues in Australia have made such measurements at the top and bottom of deep mine shafts and have claimed that there may be a real difference between their value of G and the best value from laboratory experiments. The difficulties lie in obtaining reliable samples of the density and in taking account of varying densities at greater depths.

Geophysical tests of the inverse square law

The absolute measurements of gravity described earlier, together with the comprehensive gravity surveys made over the surface of Earth, allow the mean value of gravity over the Earth to be estimated to about one part in 10^6. The techniques of space research also have given the mean value of the radius of the Earth and the distances of artificial satellites to the same precision; thus, it has been possible to compare the value of gravity on the Earth with that acting on an artificial satellite. Agreement to about one part in 10^6 shows that, over distances from the surface

of the Earth to close satellite orbits, the inverse square law is followed.

It seems at present that all the most reliable experiments and observations reveal no deviation from the inverse square law, although there should possibly be further investigations on the geophysical scale.

The principle of equivalence. Experiments with ordinary pendulums test the principle of equivalence to no better than about one part in 10^5. Eötvös obtained much better discrimination with a torsion balance. His tests depended on comparing gravitational forces with inertial forces for masses of different composition. Eötvös set up a torsion balance to compare, for each of two masses, the gravitational attraction of the Earth with the inertial forces due to the rotation of the Earth about its polar axis. His arrangement of the masses was not optimal, and he did not have the sensitive electronic means of control and reading that are now available. Nonetheless, Eötvös found that the weak equivalence principle (see above) was satisfied to within one part in 10^9 for a number of very different chemicals, some of which were quite exotic. His results were later confirmed by the Hungarian physicist J. Renner. Renner's work has been analyzed recently in great detail because of the suggestion that it could provide evidence for a new force. It seems that the uncertainties of the experiments hardly allow such analyses, however.

Eötvös suggested that the attraction of the Sun upon test masses could be compared with the inertial forces of the Earth's orbital motion about the Sun. He performed some experiments, verifying equivalence with an accuracy similar to that which he had obtained with his terrestrial experiments. The solar scheme has substantial experimental advantages, and the American physicist Robert H. Dicke and his colleagues, in a careful series of observations in the 1960s (employing up-to-date methods of servo control and observation), found that the weak equivalence principle held to about one part in 10^{11} for the attraction of the Sun on gold and aluminum (Figure 2). A later experiment, with very different experimental arrangements, by the Russian researcher Vladimir Braginski, gave a limit of about one part in 10^{12} for platinum and aluminum.

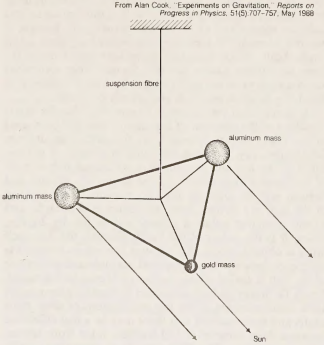


Figure 2: Experiment comparing the inertial and gravitational accelerations of gold and aluminum attracted by the Sun.

Measurements of the gravitational constant

The constant of gravitation. The constant of gravitation has been measured in three ways: (1) the comparison of the pull of a large natural mass with that of the Earth, (2) the measurement with a laboratory balance of the attraction of the Earth upon a test mass, and (3) the direct measurement of the force between two masses in the laboratory. The first approach was suggested by Newton: the earliest observations were made in 1774 by the British astronomer Nevil Maskelyne on the mountain of Schiehallion in Scotland. The subsequent work of Airy and more recent developments are noted above. The laboratory balance method was developed in large part by the British physicist John Henry Poynting during the late 1800s, but all the most recent work has involved the use of the torsion balance in some form or other for the direct laboratory measurement of the force between two bodies. The torsion balance was devised by Michell, who died before he could use it to measure G. Cavendish adapted Michell's design to make the first reliable measurement of G in 1798; his results were so good in fact that it is only in comparatively recent times that clearly better results have been obtained. Cavendish himself measured the change in deflection of the balance when attracting masses were moved from one side to the other of the torsion beam. The method of deflection was analyzed most thoroughly in the late 1800s by Sir Charles Vernon Boys, an English physicist, who carried it to its highest development using a delicate suspension fibre of fused silica for the pendulum.

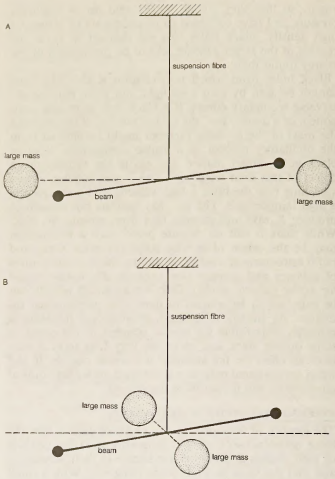


Figure 3: Principle of the determination of G from the change in period of a torsion balance with attracting masses in different positions.
From Alan Cook, "Experiments on Gravitation," *Reports on Progress in Physics*, 51(5):707–757, May 1988

The second scheme involves observing the changes in the period of oscillation of a torsion balance when attracting masses are placed close to it such that the period is shortened in one position and lengthened in another (see Figure 3). Period measurements can be made much more precisely than those of deflection, and this scheme, introduced by Carl Braun of Austria in 1897, seems to have given the best results so far. Indeed it is only recently that the value obtained by the American physicist Paul R. Heyl and his colleagues has been improved upon by Gabriel G. Luther and William R. Towler (see table).

In the third scheme, a balance with heavy attracting masses is set up near a free test balance and adjusted so that it oscillates with the same period as the test balance. The latter is then driven into resonant oscillations with an amplitude that is a measure of the constant of gravitation. The technique was first employed by J. Zahradnicek of Czechoslovakia during the 1930s and was effectively used again by C. Pontikis of France some 40 years later.

The variation of the constant of gravitation with time. The 20th-century English physicist P.A.M. Dirac, among others, suggested that the value of the constant of gravi-

Values of the Constant of Gravitation (up to 1982)

author	year	method	G (in units of 10^{-11} $m^3 s^{-2} kg^{-1}$)
H. Cavendish	1798	torsion-balance (deflection)	6.754
J.H. Poynting	1891	common-balance	6.698
C.V. Boys	1895	torsion-balance (deflection)	6.658
C. Braun	1897	torsion-balance (deflection)	6.658
	1897	torsion-balance (period)	6.658
P.R. Heyl	1930	torsion-balance (period)	6.669
J. Zahradnicek	1932	torsion-balance (resonance)	6.659
P.R. Heyl, P. Chrzanowski	1942	torsion-balance (period)	6.672
C. Pontikis	1972	torsion-balance (resonance)	6.6714
G.G. Luther and W.R. Towler	1982	torsion-balance (period)	6.6726

Possible rates of change

tation might be proportional to the age of the universe; other rates of change over time also have been proposed. The rates of change would be extremely small, one part in 10^{11} per year if the age of the universe is taken to be 10^{11} years; such a rate is entirely beyond experimental capabilities at present. There is, however, the possibility of looking for the effects of any variation upon the orbit of a celestial body, in particular, the Moon. It has been claimed from time to time that such effects may have been detected. As yet, there is no certainty.

Fundamental character of G. The constant of gravitation is plainly a fundamental quantity, since it appears to determine the large-scale structure of the entire universe. Gravity is a fundamental quantity whether it is an essentially geometric parameter as in general relativity or whether it is the strength of a field that is one aspect of a more general field of unified forces. The fact that, so far as is known, gravitation depends on no other physical factors makes it likely that the value of G reflects a basic restriction on the possibilities of physical measurement, just as special relativity is a consequence of the fact that, beyond the shortest distances, it is impossible to make separate measurements of length and time. (A.H.C.)

BIBLIOGRAPHY

General: ISAAC NEWTON, *The Mathematical Principles of Natural Philosophy*, 2 vol. (1729, reissued in 1 vol., 1975; originally published in Latin, 1 vol., 1687), often referred to as the *Principia*, is the origin of all fundamental work on gravity. STEPHEN W. HAWKING and W. ISRAEL (eds.), *Three Hundred Years of Gravitation* (1987), provides many authoritative review articles in commemoration of the tercentenary of the publication of Newton's *Principia*. CHARLES W. MISNER, KIP S. THORNE, and JOHN ARCHIBALD WHEELER, *Gravitation* (1973), is a leading work on gravitational theory. ALAN COOK, *The Motion of the Moon* (1988), discusses theories of the lunar orbit, with a chapter on applications that includes an account of gravitational studies.

Gravity around the Earth, Moon, and planets: ALAN COOK, *Physics of the Earth and Planets* (1973), includes a chapter on methods and results of gravity measurements. WOLFGANG TORGE, *Geodesy: An Introduction* (1980), contains a full and up-to-date chapter on gravity measurements. JAMES A. HAMMOND and JAMES E. FALLER, "Results of Absolute Gravity Determinations at a Number of Sites," *Journal of Geophysical Research*, 76(32):7850–7854 (Nov. 10, 1971), gives an account of absolute measurements by free fall. C. MORELLI *et al.*, *The International Gravity Standardization Net 1971 (I.G.S.N. 71)* (1974), is the result of adjusting a large number of gravity measurements over the Earth. ALAN COOK, *Interiors of the Planets* (1980), summarizes knowledge of the gravity fields of the planets and their interpretation.

Theories of gravitation and astronomical aspects: STEPHEN W. HAWKING, *A Brief History of Time: From the Big Bang to Black Holes* (1988), is a nonmathematical book by an outstanding author; see especially the chapter on black holes. CLIFFORD M. WILL, *Theory and Experiment in Gravitational Physics* (1981, reprinted 1985), is a thorough treatment. See also J.H. TAYLOR and P.M. MCCULLOCH, "Evidence for the Existence of Gravitational Radiation from Measurements of the Binary Pulsar PSR 1913+16," *Annals of the New York Academy of Sciences*, 336:442–446 (1980); and B.W. PETLEY, *The Fundamental Physical Constants and the Frontier of Measurement* (1985, reprinted 1988).

Experiments on gravitation: Experiments on aspects of gravitation are treated in the following journal articles: ALAN COOK, "Experiments on Gravitation," *Reports on Progress in Physics*, 51(5):707–757 (May 1988), a comprehensive review, with many references to early work, to the measurements of G in the table, and to recent studies of the inverse square law; P.G. ROLL, R. KROTKOV, and R.H. DICKE, "The Equivalence of Inertial and Passive Gravitational Mass," *Annals of Physics*, 26(3):442–517 (Feb. 20, 1964), a thorough account of a classical experiment; HENRY CAVENDISH, "Experiments to Determine the Density of the Earth," *Philosophical Transactions of the Royal Society of London*, 88:469–526 (June 21, 1798), the first measurement of G; GABRIEL G. LUTHER and WILLIAM R. TOWLER, "Redetermination of the Newtonian Gravitational Constant G," *Physical Review Letters*, 48(3):121–123 (Jan. 8, 1982), probably the best measurement of G up to that time; F.D. STACEY *et al.*, "Geophysics and the Law of Gravity," *Reviews of Modern Physics*, 59(1):157–174 ((Jan. 1987), a review of work on a possible difference between G as measured in the laboratory and estimated geophysically; and T.C. VAN FLANDERN, "Is the Gravitational Constant Changing?" *The Astrophysical Journal*, 248(2):813–816 (Sept. 1, 1981), a discussion of relevant observations of the Moon.

(A.H.C.)

Greece

Greece (Ellás), or the Hellenic Republic (Elliniki Dhimokratía), is the southernmost of the countries of the Balkan Peninsula. It is a land of mountains and of sea. It is difficult to be far out of range of either, a fact that has had an important influence on the country's economic and historical development. Mountains have historically restricted internal communications, but the sea has opened up wider horizons. Greece has an area of 50,949 square miles (131,957 square kilometres), of which one-fifth constitutes the Greek islands. The area of Greece is approximately the same as that of England or the U.S. state of Alabama.

The country is bordered to the west by the Ionian Sea, to the south by the Mediterranean Sea, and to the east by the Aegean Sea; only to the north and northeast does it have land borders. These run from west to east with Albania (153 miles [247 kilometres]), Macedonia (the former Yugoslav Republic of Macedonia; 159 miles [256 kilometres]), Bulgaria (295 miles [475 kilometres]), and Turkey (126 miles [203 kilometres]), totaling altogether 734 miles (1,181 kilometres).

Greece has more than 2,000 islands, of which 170 are inhabited; some of the easternmost Aegean islands lie just a few miles off the Turkish coast. Given this situation, it is no accident that Greece has always had a strong nautical tradition.

The country's capital is Athens, which has expanded rapidly in the period since World War II. The area around the capital (Attica) is now home to about one-third of the country's entire population.

A Greek legend has it that God distributed all of the available soil through a sieve and used the stones that remained to build Greece. The country's barren landscape has been a powerful factor impelling Greeks to migrate, a process that has continued for centuries until very recent times. The Greeks, like the Jews and Armenians, are a people of the diaspora; there are several million people of Greek descent in various parts of the world. *Xeniteia*, or sojourning in foreign parts, with its strong overtones of nostalgia for the faraway homeland, has been a central element in the historical experience of the Greek people.

Greece lies at the juncture of Europe, Asia, and Africa. It is heir to the heritages of classical Greece, the Byzantine Empire, and nearly four centuries of Ottoman Turkish rule. From ancient Greece the modern country inherited a sophisticated culture and a language that has been documented for almost three millennia. The language of Periclean Athens in the 5th century BC and the present-day language of the Greeks are recognizably one and the same; few languages can demonstrate such continuity. From the Byzantine Empire it has inherited Eastern Orthodox Christianity and from Ottoman rule attitudes and values that continue to be of significance, not least in shaping the country's political culture.

Greece is a country that is at once European, Balkan, and Mediterranean. It is also a country that is peculiarly burdened by its past: Greece is the only country in the world, Greek the only language, and Greeks the only people regularly prefaced by the epithet "modern." References to Greece and Greek usually denote ancient Greece and ancient Greek. Greeks, however, take great pride in their cultural heritage, and the notion of an unbroken continuity between ancient and modern Greece is an essential element in the Greek self-image.

In 1981 Greece joined the European Community (renamed the European Union in 1994). It was the first eastern European country to do so, and its heritage of Ottoman rule and Orthodox Christianity set it apart from the existing member states. The centuries of Ottoman rule have insulated the Greek lands from many of the important historical movements, such as the Renaissance, the Reformation, the Enlightenment, the French Revolution, and the Industrial Revolution, that shaped the destinies of the countries of western Europe. Membership in the European Union has been a factor in buttressing Greece's somewhat uncertain identity as a European country.

(R.R.M.C.)

This article is divided into the following sections:

Physical and human geography 179
  The land 179
    Relief
    Climate
    Drainage
    Plant and animal life
    Settlement patterns
  The people 186
    Linguistic, ethnic, and religious background
    Demography
  The economy 186
    Resources
    Agriculture, forestry, and fishing
    Industry
    Finance
    Trade
    Transportation
  Administration and social conditions 187
    Government
    Education
    Health and welfare
  Cultural life 188
    The arts
    Cultural institutions
    Daily life
    Press and broadcasting
History 189
  Greece during the Byzantine period (*c.* AD 300–*c.* 1453) 189
    Late Roman administration
    The evolution of Byzantine institutions
    Byzantine recovery
    Economy and society
    Results of the Fourth Crusade
    Population and languages
  Greece under Ottoman rule 194
    The millet system
    Disadvantages for non-Muslims
    Resistance to Ottoman rule
    Belief in divine intervention
    The role of the Orthodox church
  Transformation toward emancipation 195
    Signs of Ottoman decline
    The Phanariotes
    The mercantile middle class
    The intellectual revival
  From insurgence to independence 197
    Rigas Velestinlis
    Western encroachments
    Philikí Etaireía
    Revolt in the Peloponnese
    Factionalism in the emerging state
  Building the nation (1832–1913) 198
    Greece under Otto of Wittelsbach
    The Great Idea
    Reform, expansion, and defeat
    The early Venizélos years
  Greek history since World War I 200
    From National Schism to dictatorship
    The Metaxas regime and World War II
    Civil war and its legacy
    Restoration of democracy
Bibliography 203

Physical and human geography

THE LAND

The dominant features of the landscape

The Greek landscape is conspicuous not only for its beauty but also for its complexity and variety. Three elements dominate. The first is the sea. A glance at the map shows that the Greek mainland is indented. Arms and inlets of the sea penetrate deeply so that only a small, wedge-shaped portion of the interior mainland is more than 50 miles (80 kilometres) from the coast. The rocky headlands and peninsulas extend out to sea as island arcs and archipelagoes; indeed, islands make up roughly 18 percent of the territory of modern Greece. The southernmost part of mainland Greece, the Peleponnese Peninsula, is joined to the mainland only by the narrow isthmus at the head of the Gulf of Corinth (Korinthiakós). The country's second landscape element is its mountainousness. Roughly 80 percent of Greece is mountain terrain, much of it deeply dissected. A series of mountain chains on the Greek mainland, aligned northwest-southeast, enclose narrow parallel valleys and numerous small basins that once held lakes. With the riverine plains (most extensive toward the coast) and thin, discontinuous strips of coastal plain, these interior valleys and basins account for the third dominant feature of the Greek landscape, the lowland. Although not extensive in Greece (accounting for 20 percent of the land area), it has played an important role in the life of the country.

Relief. Three characteristics of geology and structure underlie these landscape elements. First, northeastern Greece is occupied by a stable block of old (Hercynian) hard rock. Second, younger and weaker rocks (predominantly of limestone origin) make up western and southern Greece. These were heavily folded in the Alp-building phase of the Tertiary Period (66.4 to 1.6 million years ago) when earth movements thrust the softer sediments east-northeast against the unyielding Hercynian block, producing a series of roughly parallel tectonic zones that gave rise to the mountain-and-valley relief sequence noted above. Third, both the Hercynian block and the Hellenidic (Alpine) ranges were subsequently raised and fractured by movements of the earth. These dislocations created the sunken basins of the Ionian and Aegean seas as well as the jagged edges so typical of Greece's landscape. Even today, earthquakes are all-too-frequent reminders that similar earth movements continue, particularly along the major fracture lines. Another consequence of the region's geologic instability is the widespread occurrence of marble (limestone altered by pressure and heat). Seismic disturbances are sometimes associated with volcanic explosions, notably involving the island of Thera (Santorin), which was virtually destroyed by a major eruption in the 2nd millennium BC. The vents of the Kaïméni Isles in the sea-filled explosion crater of Thera remain active. The island of Melos (Mílos), which rises to 2,464 feet (751 metres), is composed of young volcanic rocks. Thus, relief and geology provide the basis for describing the Greek landscape in terms of six major regions.

Central Greece: the Pindus Mountains. The central mountain range, the rugged Pindus (Píndhos) Mountains, forms the core of mainland Greece. Following the general northwest-southeast trend of the mountains of the Balkan Peninsula, the Pindus sweep down from the Albanian and Macedonian frontiers, creating a powerful communications barrier. Two passes (Métsovon and Mount Tímfristós) divide the range into three units: a fairly open one in the north where impervious shales and sandstones have weathered into extensive upland valleys and gently inclining hills; the Pindus proper, some 20 miles in width and predominantly limestone, in the centre; and an almost uncrossable southern zone, some 50 miles wide, deeply dissected by winding rivers and composed of a mixture of limestone, slates, and sandstones. The highest point, Mount Smólikas, 8,652 feet (2,637 metres) high, is found in the northern Pindus.

Northeastern Greece: Macedonia and Thrace. A number of topographic regions surround the main mountainous core and are often penetrated by extensions of it. The northernmost part, roughly the regions of Greek Macedonia (Makedhonía) and Thrace (Thráki), extends in a long, narrow, east-west band between the Aegean coast and the frontier with the Republic of Macedonia and Bulgaria. It consists of a series of forest-clad, crystalline mountain massifs and plateaus created by the fracturing of the old Hercynian block and separated from each other by the alluvial deposits of the five great rivers of northern Greece, the Maritsa, Néstos, Struma, Vardar, and Aliákmon rivers. The complexities of that fracturing account for the odd three-pronged shape of the Chalcidice (Khalkidhikí) Peninsula, on whose easternmost prong, Áyion

David Warren/Superstock



Church of St. Sophia at Mistra, ruined Byzantine city, on a spur of the Taÿetos Mountains overlooking olive groves in the Evrótas River valley, the Peloponnese.


SERBIA AND MONTENEGRO
Skopje
MACEDONIA
Plovdiv
BULGARIA
ADRIATIC SEA
Tirane
ALBANIA
RHODOPE MOUNTAINS
DRÁMA
XÁNTHI
ROÚPHOI
ÉVROS
Kavála
Thásos
KAVÁLA
THÁSOS ISLAND
Samothráki
SAMOTHRACE ISLAND
Alexandroúpolis
Mount Athos Monastic Republic
MOUNT ATHOS
Mount Athos 2033
Pournias Bay
Mirina
LEMNOS ISLAND
Bay of Moudhros
LESBOS
ÁYIOS EVSTRÁTIOS ISLAND
LESBOS ISLAND
AEGEAN SEA
Gulf of Kalloni
SKÝROS ISLAND
ÉVVOIA
EUBOEA ISLAND
PSARÁ ISLAND
CHIOS ISLAND
KHÍOS
Mestá
ANDROS ISLAND
TÍNOS ISLAND
MÍKONOS ISLAND
SYROS ISLAND
Tínos
Mykonos
Hermoúpolis
CYCLADES
DÉLOS ISLAND
NÁXOS ISLAND
PÁROS ISLAND
Náxos
Párga
SÉRIFOS ISLAND
SÍFNOS ISLAND
MÍLOS ISLAND
SIKINOS ISLAND
IOS ISLAND
AMORGÓS ISLAND
FOLÉGANDROS ISLAND
Thíra
ANÁFI ISLAND
THERA ISLAND
CRETE SEA
SEA OF CRETE
CRETE
DÍA ISLAND
Iráklion
Gulf of Mirabello
Réthimnon
RÉTHIMNI
IRÁKLION
Áyios Nikólaos
Ierápetra
Áyia Varvára
Phaestus
Khaniá
KHANIÁ
SAMARIA NATIONAL PARK
Khóra Sfakíon
GÁVDOS ISLAND
Kíthira
CYTHERA ISLAND
ANDIKÍTHIRA ISLAND
ATTIKÍ
Athens
Piraeus
Kallithéa
Glyfáda
Peristéri
Néa Ionía
Kifisiá
Marathónas
Mount Pentelikon 1109
Mount Parnitha 1413
Mount Hymettus 1026
Mount Penteli 1280
Kaisariani
Vouliagmeni
Lávrion
Kéa
Aegina
Sounio
Aegina Island National Park
Poros Island
Methana
Merikhas
Kithnos Island
Hydra Island
Idhra
Spétsai
Ermionídha
Mount Kandhíli 1057
Mount Ortholithi 1976
Mount Profitis Ilías 1935
Mount Kyllini 2376
Mount Chelmós 2355
Mount Aroánia 1926
Mount Erymanthos 2224
Mount Olonos 1976
Mount Skollis 1645
Mount Lykaion 1420
Mount Taygetos 2404
Mount Parnon 1935
Mount Malevós 1407
Mount Párnon 1954
Mount Kalmaxitsa 2295
Mount Vérmio 2052
Mount Voras 2524
Mount Vermion 2052
Mount Olympus 2917
Mount Ossa 1978
Mount Pelion 1657
Mount Othrys 1726
Mount Giona 2510
Mount Parnassos 2457
Mount Kithairon 1409
Mount Helicon 1748
Mount Kerketeus 2344
Mount Dirfys 1743
Mount Ochi 1398
Mount Prophet Elias 1172
Mount Kedros 1780
Mount Ida 2456
Mount Psiloritis 2456
Mount Dikti 2148
Mount Afendis Kavousi 1476
Mount Thrypti 1476
Mount Orno 1476
Mount Kofinas 1231
Mount Lasithi 1231
Mount Kastelli 1476
Mount Kadistos 1231
Mount Kofinas 1231
Mount Kofinas 1231
Mount Kofinas



MAP INDEX

Political subdivisions

Aetolia and Acarnania (Aitolia kai Akarnania) . . . . 38 30 N 21 30 E
Achaea (Akhaia) . . 38 00 N 22 00 E
Arcadia (Arkadhia) . . . . . 37 35 N 22 15 E
Argolis . . . . . 37 40 N 22 50 E
Árta . . . . . 39 10 N 21 00 E
Attiki . . . . . 38 05 N 23 30 E
Áyion Óros, see Mount Athos
Boetia (Voiotia) . . . 38 20 N 23 00 E
Corfu (Kérkira) . . . . 39 40 N 19 45 E
Corinth (Korinthia) . . . . . 37 55 N 22 40 E
Cyclades (Kikládhes) . . . . . 37 25 N 24 55 E
Dhodhekánisos . . . 36 50 N 27 05 E
Dráma . . . . . 41 15 N 24 15 E
Evritania . . . . . 39 00 N 21 40 E
Évros . . . . . 41 00 N 26 00 E
Évvoia . . . . . 38 30 N 24 00 E
Flórina (Phlorina) . . . 40 45 N 21 25 E
Fokis (Phocis) . . . . 38 30 N 22 15 E
Fthiótis . . . . . 38 50 N 22 25 E
Grevená . . . . . 40 05 N 21 25 E
Ilia . . . . . 37 45 N 21 35 E
Imathia . . . . . 40 30 N 22 15 E
Ioánnina . . . . . 39 45 N 20 40 E
Iráklion . . . . . 35 10 N 25 10 E
Kardhitsa . . . . . 39 20 N 21 45 E
Kastoría . . . . . 40 30 N 21 10 E
Kavála . . . . . 41 00 N 24 30 E
Kefallinía . . . . . 38 15 N 20 30 E
Kérkira, see Corfu
Khalkidhiki . . . . . 40 25 N 23 30 E
Khaniá . . . . . 35 20 N 24 00 E
Khíos . . . . . 38 25 N 26 00 E
Kikládhes, see Cyclades
Kilkís . . . . . 41 00 N 22 40 E
Korinthia, see Corinth
Kozáni . . . . . 40 15 N 21 35 E
Laconia (Lakonia) . . 37 00 N 22 35 E
Lárisa . . . . . 39 30 N 22 30 E
Lasithi . . . . . 35 05 N 25 50 E
Lesbos (Lésvos) . . 39 10 N 26 20 E
Levkás . . . . . 38 45 N 20 40 E
Magnisia . . . . . 39 15 N 22 45 E
Messenia (Messinía) . . . 37 15 N 21 50 E
Mount Athos (Áyion Óros) . . . . 40 15 N 24 15 E
Pélla . . . . . 40 50 N 22 15 E
Phlorina, see Flórina
Phocis, see Fokis
Pieria . . . . . 40 15 N 22 25 E
Préveza . . . . . 39 10 N 20 40 E
Rethímni . . . . . 35 15 N 24 35 E
Rodhópi . . . . . 41 05 N 25 30 E
Sámos . . . . . 37 45 N 26 45 E
Sérrai . . . . . 41 10 N 23 30 E
Thesprotía . . . . . 39 30 N 20 20 E
Thessaloníki . . . . 40 40 N 23 00 E
Trikala . . . . . 39 40 N 21 30 E
Voiotia, see Boetia
Xánthi . . . . . 41 10 N 24 50 E
Zákinthos . . . . . 37 45 N 20 45 E

Cities and towns

Agrínion . . . . . 38 38 N 21 25 E
Áyion (Egion) . . . . 38 15 N 22 05 E
Alexandroúpolis (Alexandhroúpolis) . . . . . 40 51 N 25 52 E
Almirós . . . . . 39 11 N 22 46 E
Amalías . . . . . 37 48 N 21 21 E
Amphissa (Amfissa) . . . . . 38 32 N 22 23 E
Ándros . . . . . 37 50 N 24 56 E
Aradípolis . . . . . 36 40 N 22 23 E
Árgos . . . . . 37 38 N 22 44 E
Árgos Orestikón . . . 40 28 N 21 16 E
Argostóli on . . . . . 38 11 N 20 29 E
Arkhángelos . . . . . 36 12 N 28 08 E
Árta . . . . . 39 09 N 20 59 E
Athens (Athinai) . . . 37 59 N 23 44 E
Áyia Varvára . . . . 35 08 N 25 00 E
Áyios Kírikos . . . . 37 35 N 26 14 E
Áyios Nikólaos . . . 35 11 N 25 43 E
Candia, see Iráklion
Canea, see Khaniá
Chalcis, see Khalkis
Chios, see Khíos
Corinth (Kórinthos) . . . . 37 56 N 22 56 E
Dhiavatá . . . . . 40 41 N 22 51 E
Dhimitsána . . . . . 37 36 N 22 03 E
Dhornokós . . . . . 39 08 N 22 18 E
Dráma . . . . . 41 09 N 24 09 E
Édhessa (Edessa) . . . . . 40 48 N 22 03 E
Egion, see Áyion
Elassón . . . . . 39 54 N 22 11 E
Ermoúpolis, see Hermoúpolis
Evzonion . . . . . 40 40 N 22 55 E
Fársala . . . . . 39 18 N 22 23 E
Filippoi, see Philippi
Flórina . . . . . 40 47 N 21 24 E
Gargaliánoi . . . . . 37 04 N 21 38 E
Githíon, see Ýthion
Giúfadha (Glyphádha) . . . . . 37 52 N 23 45 E
Grevená . . . . . 40 05 N 21 25 E
Herákleion, see Iráklion
Hermoúpolis (Ermoúpolis) . . . . 37 27 N 24 56 E
Idhra . . . . . 37 21 N 23 28 E
Ierápetra . . . . . 35 01 N 25 45 E
Igoumenítsa . . . . . 39 30 N 20 16 E
Ioánnina (Yannina) . . . . . 39 40 N 20 50 E
Ios . . . . . 36 44 N 25 17 E
Iráklion (Candia or Herákleion) . . . . . 35 20 N 25 08 E
Istiaía . . . . . 38 57 N 23 09 E
Itéa . . . . . 38 26 N 22 25 E
Kalabáka . . . . . 39 42 N 21 38 E
Kalámai (Kalamáta) . . . . . 37 02 N 22 07 E
Kalamariá . . . . . 40 36 N 22 58 E
Kalávrita . . . . . 38 02 N 22 07 E
Kálimnos . . . . . 36 57 N 26 59 E
Kallithéa . . . . . 37 57 N 23 42 E
Kardhitsa . . . . . 39 22 N 21 55 E
Karlsá . . . . . 40 15 N 24 15 E
Kárystos . . . . . 38 01 N 24 25 E
Karpenísion . . . . . 38 55 N 21 47 E
Kastoría . . . . . 40 31 N 21 16 E
Katerini . . . . . 40 16 N 22 30 E
Káto Akhaia . . . . . 38 09 N 21 33 E
Kavála (Kaválla or Neapolis) . . . . . 40 56 N 24 25 E
Kéa . . . . . 37 38 N 24 21 E
Kérkira . . . . . 39 36 N 19 55 E
Khalkidhón (Nea Khalkidhón) . . . . 40 44 N 22 36 E
Khalkís (Chalcis) . . 38 28 N 23 36 E
Khaniá (Canea) . . . 35 31 N 24 02 E
Khíos (Chios) . . . . 38 22 N 26 08 E
Khóra Sfakíon . . . . 35 12 N 24 09 E
Kifisiá (Kifissia) . . . 38 04 N 23 49 E
Kilkís . . . . . 41 00 N 22 52 E
Kími . . . . . 38 38 N 24 06 E
Kiparissía . . . . . 37 15 N 21 40 E
Kíthira . . . . . 36 09 N 22 59 E
Komotiní . . . . . 41 07 N 25 24 E
Kórinthos, see Corinth
Koróni . . . . . 37 52 N 24 03 E
Kos . . . . . 36 53 N 27 18 E
Kozáni . . . . . 40 18 N 21 47 E
Kranídhion . . . . . 37 23 N 23 09 E
Lágos . . . . . 41 01 N 25 07 E
Lamía . . . . . 38 54 N 22 26 E
Lárisa (Lárissa) . . . 39 38 N 22 25 E
Lávrion (Laurium) . . 37 43 N 24 03 E
Leondárion . . . . . 39 11 N 22 08 E
Leonídhion . . . . . 37 10 N 22 52 E
Levádhia . . . . . 38 26 E 22 53 E

Levkás ..... 38 50 N 20 42 E
Lindhos (Lindos) .. 36 06 N 28 04 E
Megalópolis ..... 37 24 N 22 08 E
Mégara ..... 38 00 N 23 21 E
Mérikhas ..... 37 23 N 24 24 E
Mesolóngion (Missolonghi) .... 38 22 N 21 26 E
Mestá ..... 38 16 N 25 55 E
Métsovon ..... 39 46 N 21 11 E
Mikonos ..... 37 27 N 25 20 E
Mitilíni (Mytilene) .. 39 06 N 26 33 E
Monemvasía ..... 36 41 N 23 03 E
Náousa ..... 40 38 N 22 04 E
Návpaktos ..... 38 24 N 21 50 E
Návplion (Nauplia) ..... 37 34 N 22 48 E
Náxos ..... 37 06 N 25 23 E
Néa Ionía ..... 38 02 N 23 45 E
Néa Khalkidhón, see Khalkidhón
Neápolis ..... 36 31 N 23 04 E
Neápolis ..... 40 19 N 21 23 E
Neapolis, see Kavála
Néon Karlovásion .. 37 47 N 26 42 E
Níkea (Nikaia) .... 37 58 N 23 39 E
Orestiás ..... 41 30 N 26 31 E
Párga ..... 39 17 N 20 24 E
Páros ..... 37 05 N 25 09 E
Pátrai ..... 38 15 N 21 44 E
Pérama ..... 39 42 N 20 51 E
Peristérion ..... 39 01 N 23 42 E
Philippi (Filippoi) .. 41 02 N 24 20 E
Pílos ..... 36 55 N 21 42 E
Piraeus (Piraiévs) .. 37 57 N 23 38 E
Políyiros ..... 40 23 N 23 27 E
Préveza ..... 38 57 N 20 45 E
Ptolemaís ..... 40 31 N 21 41 E
Pyrgos (Pírgos) .... 37 41 N 21 27 E
Réthimnon ..... 35 22 N 24 28 E
Ródhos (Rhodes) . 36 26 N 28 13 E
Salamis (Salamís) . 37 58 N 23 29 E
Salonika, see Thessaloniki
Sámos ..... 37 45 N 26 58 E
Samothráki ..... 40 29 N 25 31 E
Sérrai ..... 41 05 N 23 33 E
Siátista ..... 40 16 N 21 33 E
Sidhirókastron .... 41 14 N 23 24 E
Sikéai (Sikiés) .... 37 31 N 21 53 E
Souflion ..... 41 12 N 26 18 E
Sparta (Spárti) .... 37 05 N 22 26 E
Stavroúpolis .... 41 12 N 24 42 E
Stení Dhirfios .... 38 35 N 23 50 E
Tanágra ..... 38 19 N 23 32 E
Thásos ..... 40 47 N 24 43 E
Thebes (Thívai) .... 38 19 N 23 19 E
Thessaloniki (Salonika) ..... 40 38 N 22 56 E
Thira ..... 36 25 N 25 26 E
Tínos ..... 37 32 N 25 10 E
Trikala ..... 39 33 N 21 46 E
Tripolis ..... 37 31 N 22 22 E
Valestínon ..... 39 23 N 22 45 E
Véroia ..... 40 31 N 22 12 E
Vólos ..... 39 22 N 22 57 E
Voúla ..... 37 51 N 23 46 E
Xánthi ..... 41 08 N 24 53 E
Xilókastron ..... 38 05 N 22 38 E
Yannina, see Ioánnina
Yiannitsá ..... 40 48 N 22 25 E
Yithion (Githion) .. 36 45 N 22 34 E
Zákinthos ..... 37 47 N 20 54 E

Physical features and points of interest

Achelous (Akhelóös), *river* .. 38 20 N 21 06 E
Aegean Sea ..... 39 00 N 25 00 E
Aegina, see Saronic Gulf
Aegina (Aíyina) Island ..... 37 44 N 23 30 E
Ainos National Park ..... 38 07 N 20 40 E
Alákmon, *river* .. 40 30 N 22 40 E
Alónnisos Island .. 39 13 N 23 55 E
Alpheus (Alfiós), *river* .... 37 37 N 21 27 E
Amorgós Island .. 36 50 N 25 59 E
Amvrakikós, see Árta, Gulf of
Anáfi Island ..... 36 22 N 25 47 E
Andros (Ándros) Island ..... 37 50 N 24 50 E
Arachthos (Árakhthos), *river* ..... 39 01 N 21 03 E
Argolis (Argolikós), Gulf of ..... 37 20 N 22 55 E
Árta (Amvrakikós), Gulf of ..... 39 00 N 21 00 E
Astipálaia Island .. 36 35 N 26 20 E
Athos (Áthos or Áyion), Mount ..... 40 10 N 24 20 E
Axiós, see Vardar
Áyios Evstrátios Island ..... 39 31 N 25 00 E
Balkan Peninsula .. 41 30 N 23 00 E
Candia, see Crete, Sea of
Castelrosso, see Kastellórizon Island
Catalonia (Cephallenia or Kefallinía) Island ..... 38 15 N 20 35 E
Chalcidice (Khalkidhikí) Peninsula ..... 40 25 N 23 27 E
Chios (Khíos) Island ..... 38 22 N 26 00 E
Corfu (Corcyra or Kérkira) Island ..... 39 40 N 19 45 E
Corinth (Korinthiakós), Gulf of ..... 38 12 N 22 30 E
Corinth (Korinthou) Canal ..... 37 57 N 22 58 E
Cos (Kos) Island .. 36 50 N 27 10 E
Crete (Kríti), *island* ..... 35 15 N 24 45 E
Crete (Candia or Kritikón), Sea of ..... 36 00 N 25 00 E
Cyclades (Kikládhes), *islands* ..... 37 00 N 25 10 E
Cythera (Kithira) Island ..... 36 15 N 23 00 E
Delos (Dhílos) Island ..... 37 24 N 25 16 E
Delphi, *historical site* ... 38 30 N 22 31 E
Dhikti (Lasithi) Mountains ..... 35 08 N 25 28 E
Dhirfis, Mount .... 38 38 N 23 50 E
Dodecanese (Dhodhekánisos or Sporades), *islands* ..... 36 00 N 27 00 E
Doran (Dhóiranis), Lake ..... 41 13 N 22 44 E
Drámas (Drama) Plain ..... 41 05 N 24 06 E
Epidaurus, *historical site* ... 37 38 N 23 09 E
Euboea (Évvoia) Island ..... 38 30 N 24 00 E
Évros, see Maritsa
Evrótas, *river* ..... 36 48 N 22 41 E
Evvoikos, see Euboea, Gulf of
Folégandros Island ..... 36 38 N 24 54 E
Helicon (Elikón), Mount ..... 38 20 N 22 50 E
Hydra (Idhra) Island ..... 37 20 N 23 30 E
Idhi (Psiloritis), Mount ..... 35 15 N 24 45 E
Ikaria Island ..... 37 35 N 26 10 E
Ionian (Iónioi) Islands ..... 38 30 N 20 30 E
Ionian Sea ..... 39 00 N 19 00 E
Ios Island ..... 36 42 N 25 20 E
Ithaca (Itháki) Island ..... 38 24 N 20 40 E
Kaimaktchalán (Kaimakčalan), Mount ..... 40 58 N 21 48 E
Kalamáta, see Messenia, Gulf of
Kálimnos Island ... 37 00 N 27 00 E
Kalloni (Kallonís), Gulf of ..... 39 07 N 26 08 E
Kamvoúnia Mountains ..... 40 00 N 21 52 E
Kárpathos (Scarpanto) Island ..... 35 40 N 27 10 E
Kásos Island ..... 35 25 N 26 55 E
Kassándra (Toronaios), Gulf of ..... 40 06 N 23 30 E
Kassándra Peninsula ..... 40 00 N 23 30 E
Kastellórizon (Megiste) Island ..... 36 08 N 29 34 E
Kastorias, Lake ... 40 31 N 21 18 E
Kéa (Tziá) Island .. 37 37 N 24 20 E
Kefallinia, see Cephalonia Island
Kérkira, see Corfu Island
Khálki Island ..... 36 14 N 27 36 E
Khalkidhikí, see Chalcidice Peninsula
Khaniá (Khanion), Gulf of ..... 35 35 N 23 50 E
Khíos, see Chios Island
Kikládhes, see Cyclades
Killini (Kyllini), Mount ..... 37 55 N 22 26 E
Kiparissia (Kiparissiakós), Gulf of ..... 37 30 N 21 25 E
Kissavos, see Óssa, Mount
Kithairón, Mount .. 38 12 N 23 15 E
Kíthira, see Cythera
Kithnos Island .... 37 23 N 24 25 E
Knossos, *historical site* ... 35 18 N 25 10 E
Korinthiakós, see Corinth, Gulf of
Korinthou, see Corinth Canal
Korónia, Lake .... 40 41 N 23 09 E
Kos, see Cos Island
Kritikón, see Crete, Sea of
Kyllini, see Killini Mount
Laconia (Lakonikós), Gulf of ..... 36 35 N 22 40 E
Lasithi, see Dhikti Mountains
Lemnos (Límnos) Island ..... 39 55 N 25 15 E
Léros Island ..... 37 08 N 26 50 E
Lesbos (Lésvos or Mitilíni) Island ..... 39 10 N 26 32 E
Leukas (Levkás) Island ..... 38 43 N 20 38 E
Levká Mountains .. 35 20 N 24 00 E
Macedonia, *region* ..... 41 00 N 23 00 E
Maina (Mani) Peninsula ..... 36 35 N 22 25 E
Maritsa (Évros), *river* ..... 40 52 N 26 12 E
Megáli Prespa, see Prespa, Lake
Megiste, see Kastellórizon Island
Melos (Milos) Island ..... 36 41 N 24 25 E
Messenia (Kalamáta or Messiniakós), Gulf of ..... 36 45 N 22 10 E
Mesta, see Néstos
Mikonos Island ... 37 27 N 25 23 E
Mikrá Préspa, Lake ..... 40 45 N 21 06 E
Milos, see Melos Island
Milos ..... 36 45 N 24 26 E
Mirabéllo (Mirabéllou), Gulf of ..... 35 14 N 25 47 E
Mithimna ..... 39 22 N 26 10 E
Mitilíni, see Lesbos Island
Moúdhrou, Bay of ..... 39 50 N 25 15 E
Mount Athos Monastic Republic ..... 40 15 N 24 15 E
Mount Olíti National Park ... 38 28 N 22 16 E
Mount Olympus National Park ... 40 05 N 22 21 E
Mycenae, *historical site* ... 37 43 N 22 45 E
Náxos Island ..... 37 02 N 25 35 E
Néstos (Mesta), *river* ..... 40 51 N 24 44 E
Nísiros Island .... 36 35 N 27 10 E
Northern Sporades (Vórioi Sporádhes) ..... 39 15 N 23 55 E
Oíti, Mount ..... 38 49 N 22 17 E
Ókhi, Mount ..... 38 04 N 24 28 E
Olympus (Ólimbos), Mount ..... 40 05 N 22 21 E
Ossa (Óssa), Mount ..... 39 49 N 22 40 E
Óthris, Mount .... 39 02 N 22 37 E
Pagasaí (Pagasitikós), Gulf of ..... 39 15 N 23 00 E
Pangaion (Pangaíon), Mount ..... 40 55 N 24 05 E
Parnassus (Parnassós), Mount ..... 38 32 N 22 37 E
Parnassus National Park .... 38 32 N 22 37 E
Párnis (Párnitha) National Park ... 38 10 N 23 40 E
Páros Island ..... 37 06 N 25 12 E
Pátmos Island .... 37 20 N 26 33 E
Patras, Gulf of ..... 38 15 N 21 30 E
Paxos (Paxoi) Island ..... 39 12 N 20 10 E
Pelion (Pílion), Mount ..... 39 24 N 23 05 E
Peloponnese (Peloponnesus or Pelopónnisos), *region* ..... 37 30 N 22 00 E
Pentelikus (Pendéli or Pentelikon), Mount ..... 38 04 N 23 53 E
Phaestus (Phaestos), *historical site* ... 35 01 N 24 48 E
Pindus (Píndhos or Píndhou) Mountains ..... 39 45 N 21 30 E
Pindus National Park ..... 39 54 N 21 14 E
Piniós (Peneus), *river* ..... 39 53 N 22 44 E
Póros Island ..... 37 30 N 23 31 E
Pourniás (Pourniá) Bay .... 40 00 N 25 19 E

Prespa (Megáli Préspa), Lake ... 40 55 N 21 00 E
Prespa National Park ..... 40 45 N 21 04 E
Psará Island ..... 38 35 N 25 37 E
Psiloritis, see Idhi, Mount
Pylos, *historical site* ..... 36 56 N 21 40 E
Rhodes (Ródhos), *island* ..... 36 10 N 28 00 E
Rhodope (Rodhópis) Mountains ..... 41 30 N 24 30 E
Samaria National Park ..... 35 17 N 24 00 E
Sámos Island .... 37 45 N 26 48 E
Samothrace (Samothráki) Island ..... 40 27 N 25 35 E
Santorin, see Thera Island
Saronic (Aegina or Saronikós) Gulf ..... 37 45 N 23 30 E
Scarpanto, see Kárpathos Island
Sérifos Island .... 37 10 N 24 30 E
Sikinos Island .... 36 41 N 25 07 E
Símí Island ..... 36 35 N 27 50 E
Singitic (Singitikós) Gulf . 40 12 N 23 56 E
Siphnus (Siphnos or Sífnos) Island ... 37 00 N 24 40 E
Skópelos Island ... 39 10 N 23 40 E
Skýros (Skíros) Island ..... 38 53 N 24 32 E
Smólikas, Mount .. 40 06 N 20 55 E
Soúnio National Park ..... 37 45 N 24 00 E
Sporades, see Dodecanese
Struma (Strimón), *river* ..... 40 47 N 23 51 E
Syros (Síros) Island ..... 37 26 N 24 55 E
Taíyetos Mountains ..... 37 06 N 22 18 E
Tempe (Témbi), Vale of ..... 39 53 N 22 35 E
Thasos (Thásos) Island ..... 40 40 N 24 40 E
Thera (Santorin or Thíra) Island .... 36 24 N 25 26 E
Thérmai (Thermaïkós), Gulf of ..... 40 23 N 22 47 E
Thermopylae (Thermopílai) Pass ..... 38 48 N 22 32 E
Thiamis, *river* .... 39 32 N 20 08 E
Thrace (Thráki), *region* ..... 41 20 N 26 45 E
Thriptis (Thríti) Mountains ..... 35 05 N 25 52 E
Tílos Island ..... 36 25 N 27 25 E
Tínos Island .... 37 35 N 25 10 E
Trikhonis (Trichonis), Lake . 38 34 N 21 30 E
Tzia, see Kéa Island
Vardar (Axiós), *river* ..... 40 35 N 22 50 E
Vegorritis, Lake ... 40 45 N 21 48 E
Vérmion Mountains ..... 40 30 N 22 00 E
Vikos-Aóös National Park .... 39 58 N 20 50 E
Vistonis, Lake .... 41 03 N 25 07 E
Vólvi, Lake ..... 40 41 N 23 28 E
Zacynthus (Zákinthos) Island ..... 37 47 N 20 47 E

(Holy) Mountain, is located Mount Athos (Áthos), the famous site of Greek Orthodox monastic communities. Along and beyond the Bulgarian border rise the Rhodope (Rodhópis) Mountains, composed mainly of sharp-edged and frequently sloping plateaus, often rising more than 7,000 feet (1,800 metres) and reaching 7,287 feet (2,212 metres) at Mount Órvilos. The Maritsa (Évros) River in its low-lying, marshy valley marks the Turkish border. From here to the lower Struma (Strimón) River extends a succession of plains, often swampy (like the deltaic plain of the lower Néstos), some of which have been turned into fertile agricultural land (like the former Lake Akhinós). Inland there are basins of structural origin (such as the Dhrámas Plain). The lakes of Korónia and Vólvi, which separate the Chalcidice Peninsula from the rest of the coastal region, also occupy structural depressions. Farther west the large plain drained by the Vardar (Axiós) and lower Aliákmon rivers is being continually extended as the river deltas push out into the Gulf of Thérmai (Thermaïkós). The forested Vérmion (Virmion) Mountains and, beyond them, the barren inland basins around Lakes Vegorritis and Kastorías mark the boundary with the Pindus Mountains.

Eastern Greece: Thessaly and Attikí. This region epitomizes the physical geography of Greece. To the west are the massive limestones so characteristic of northern and western Greece, while to the east the peninsula of Attikí (Attica) represents the western margin of the old (Hercynian) crystalline rocks of the Aegean shores. Essentially an upland area, its relief is articulated by four northwest-southeast-trending spurs thrusting out from the main Pindus mass. A number of distinctive basins and plains lie amid these upland ribs. The northernmost, a rather broken spur called the Kamvoúnia Mountains, runs along the coast of the Gulf of Thérmai and continues south to form the peninsula bounding one side of the Vólou Bay. Among its peaks are Mount Olympus (Ólimbos)—the mythical seat of the gods, whose often cloud-topped summit rises to 9,570 feet (2,917 metres), the highest point in Greece—and the equally fine peaks of Mount Óssa and Mount Pelion (Pílion). The next spur on the west is the Óthris mountain range, which continues across the narrow Oreón (Oreón) Channel in the northern sector of the long, narrow island of Euboea (Évvoia). Between the two spurs lie the ancient basins (formerly the site of lakes) of Thessaly (Thessalía), Tríkala (Trikkala), and Lárisa, drained by the Piniós. Just to their south the basin of Almirós, of similar origin, lies around Vólou Bay.

To the southwest, the third spur leaving the Pindus is that of the Oíti, continued in the Ókhi Mountains of southern Euboea. Just before the Oíti reaches the sea, near the head of the Gulf of Maliakós, is the pass of Thermopylae (Thermopílai), scene of the famous battle of antiquity. The last (and perhaps the most important) of the four spurs thrusting down into eastern Greece is the one that curves away to the southeast through the twin-peaked mass of Mount Parnassus (Parnassós). This mountain rises to 8,061 feet (2,457 metres) and was held to be the home of the Muses. The view from its summit at sunrise, with a broad expanse of the heart of Greece gradually unfolding, is regarded as one of the finest in the world. The range continues as the backbone of the peninsula lying between the Gulf of Euboea and the Gulf of Corinth, and it reaches as far as Mount Párnis, just to the north of Athens. To its north lie the plains of Phocis (Fokís) and Boeotia (Voiotía), and around its southern tip lie the depressions of Attikí, hotter and more arid but with a strategic importance that helps to explain the rise of Athens.

Mount Parnassus and its regional context

Southern Greece: the Peloponnese. The entire southern portion of mainland Greece forms a peninsula lying to the south of the Gulf of Corinth. Technically, this region, the Peloponnese, or Pelopónnisos, also known as the Morea, is now an island, for the 3.9-mile Corinth Canal cuts across the narrow neck of land formerly separating the Gulf of Corinth from that of Aegina (Aíyina). The Peloponnese consists of an oval-shaped mountain mass with peaks rising to 7,800 feet and four peninsular prongs that point southward toward the island of Crete. At its heart are the arid limestone plateaus of Arcadia (Arkadhía), where streams disappear underground into the soluble rock and from which the barren upland of the Taíyetos Mountains (7,800 feet) extends southward to form the backbone of one of the southern peninsulas. A thin fringe of fertile coastal plain in the north and west, together with the larger alluvial depressions forming the Gulfs of Laconia (Lakonikós), Messenia (Messiniakós), and Árgolis, surrounds this mountainous core. The coast is indented and offers some fine harbours.

Western Greece: Epirus and Arkananía. The distinctiveness of the western side of the Greek mainland (consisting of Epirus [Ípiros] and Arkananía) north of the Gulf of Corinth to the Albanian frontier and the offshore Ionian (Iónioi) Islands is enhanced by the fact that the barrier effect of the Pindus and the ameliorating climatic influences from the west that result in a quite different landscape from that of the rest of Greece have exaggerated the historic isolation from the other areas of mainland Greece. Fertile basins are not well developed, constricted as they are by the parallel ranges of the coastal mountains. The mountain regions themselves, however, are adequately supplied with rainfall. The flat, alluvial plain of Árta, built up from detritus brought down by the Arachthos (Árakhthos) River, has become, with irrigation, a fertile agricultural region.

The islands of Greece. The Ionian Islands off the western coast of Greece structurally resemble the folded mountains of Epirus. Of the seven main islands, Corfu (Kérkira), opposite the Albanian frontier, is the northernmost. It is fertile and amply endowed with well-watered lowland. The other islands, Paxos (Paxoí), Leukas (Levkás), Skorpiós, Ithaca (Itháki), Cephalonia (Kefallinía), and Zacynthus (Zákinthos), lie farther south. Lack of rainfall accentuates their gaunt, broken limestone relief, although Leukas and Zacynthus have sheltered eastern plains. The Aegean Islands, also exhibiting the characteristic landforms of the mainland, are situated in distinct clusters in the Aegean Sea, east of the Greek mainland.

Regional clusters of islands

In the north, off Thrace, lie Thásos (an oval block of ancient mineral rocks similar in composition to neighbouring blocks on the mainland) and harbourless Samothrace (Samothráki), an island of volcanic origin. Lemnos (Lím-



Sunlight on the whitewashed walls of Thíra, chief town of Thera, the southernmost island of the Cyclades group of the Aegean Islands.
Ivan Massar—Black Star

nos), situated midway between Asia Minor and Áyion Mountain peninsula, is almost cut in two by the northern Pourniás Bay and the deep southern harbour afforded by the Bay of Moúdhrou.

To the southeast the rocky but sheltered islands of Lesbos (Lésvos), Chios (Khíos), and Sámos lie close to the Turkish coast and are extensions of peninsulas on the coast of Asia Minor. Across the central Aegean, near northern Euboea, lie the Northern Sporades (Vórioi Sporádhes), or "Scattered Islands"; their crystalline rocks are similar to those of the Greek mainland. Farther south, in the heart of the Aegean, lie the Cyclades (Kikládhes), "Islands in a Circle." These roughly centre on Delos (Dhílos) and represent the tips of drowned mountain ridges continuing the structural trends of Euboea and the region around Athens.

Between the Cyclades and the Turkish coast, the Dodecanese (Dhodhekánisos) group, with Rhodes (Ródhos) the largest of a dozen major islands, has a varied geologic structure ranging from the gray limestones of Kálimnos, Símі, and Khálki to the complete ancient volcanic cone that forms Nísiros.

Crete

Finally, the long narrow shape of Crete (Kríti) stands to the south at the entrance of the Aegean. By far the largest of the Aegean Islands and the fifth largest island in the Mediterranean (3,190 square miles), Crete is geologically linked to the south and west of mainland Greece. Its rugged, deeply ravined, asymmetrical limestone massif, falling steeply to the south, is so divided as to resemble four separate islands when seen from a distance: the westernmost Lévka ("White") Mountains; the central Ídhi (or Psiloritis) Mountains, with Crete's highest point, the summit of Mount Ídhi, Stavros, 8,058 feet (2,456 metres) high; the east-central Dhíkti Mountains; and the far eastern Thriptís (Thrifti) Mountains. Another range, the Asterousia (or Kófinos) Mountains, runs along the south-central coast between the Mesará Plain and the Libyan Sea. Of Crete's 650 miles (1,046 kilometres) of rocky coastline, it is the more gradual slope on the northern side of the island that provides several natural harbours and coastal plains. (J.S.Bo./C.D.S.)

Climate. The basically Mediterranean climate of Greece is subject to a number of regional and local variations occasioned by the country's physical diversity. In winter the belt of low-pressure disturbances moving in from the North Atlantic shifts southward, bringing with it warm, moist, westerly winds. Squalls and spells of rain ruffle the Aegean, but sunshine often breaks through the clouds. As the low-pressure areas enter the Aegean region, they may draw in cold air from those eastern regions of the Balkans that, sheltered by the Dinaric mountain system from western influences, are open to climatic extremes emanating from the heart of Eurasia. This icy wind is known as the boreas. Partly as a result, Thessaloníki (Salonika) has an average January temperature of 43° F (6° C), while Athens has 50° F (10° C) and Iráklion (Hérakleion) 54° F (12° C). *Shílok*, or warm winds, are similarly drawn in from the south. The western influences bring plentiful rain to the Ionian coast and the mountains behind it; winter rain also starts early, and snow lingers into spring. At Corfu, January temperatures average 50° F (10° C), and the island's average annual rainfall is 52 inches (1,320 millimetres), compared with the total on Crete of 25 inches and the total at Athens of 16 inches. On Crete, snow is almost permanent on the highest peaks.

The boreas

In summer, when the low-pressure belt swings away again, the climate is hot and dry almost everywhere, with the average July sea-level temperature approaching 80° F (27° C), although heat waves can push the temperature up over the 100° F (38° C) mark for a day or so. Topography is again a modifying factor: the interior northern mountains continue to experience some rainfall, while all along the winding coast the afternoon heat is eased slightly by sea breezes. In other regions, such as Crete, the hot, dry summers are accentuated by the parching *neltemi*, or Etesian winds, which become drier and drier as they are drawn southward.

In all seasons—perhaps especially in summer—the quality of light is one of Greece's most appealing attractions. However, atmospheric pollution has become a serious problem in the cities, notably Athens, and a hazard to the ancient monuments.

Drainage. The main rivers of Greece share several characteristics: in their upper courses most flow in broad, gently sloping valleys; in their middle courses they plunge from intermontane basin to basin through narrow, often spectacular gorges; in their lower courses they meander across the coastal plain to reach the sea in marshy, ever-growing deltas. Most rivers are very short. In limestone districts a generally permeable surface with sinkholes (*katavóthra*) leading to underground channels complicates the drainage network. In all regions river regimes are erratic, unsuitable for navigation, and of limited usefulness for irrigation. The Vardar, Struma, and Néstos, which cross Greek Macedonia and Thrace to enter the northern Aegean, are the major rivers, but only because they drain large regions beyond the Greek frontier. Also in the northeast are the eastward-flowing Aliákmon and Piniós (Peneus). In the Peloponnese, only the Evrótas is noteworthy.

Plant and animal life. As in other Balkan countries, the vegetation of Greece is open to influences from several

major biogeographic zones, with the major Mediterranean and western Asian elements supplemented by plants and animals from the central European interior. Add to this the climatic effects of altitude, the contrast between north and south, and the role of local relief, together with the ubiquitous human factor, the result of some eight or nine millennia of settlement and land use, and it is not difficult to appreciate either the subtlety or the complexity of the vegetation mosaic. Degraded plant associations (reduced in variety and height of species and density of plant cover) and soil erosion are commonplace.

On the mountain flanks, and in the north generally, the central European types of vegetation prevail. In central and southern regions and in narrow belts along the valleys of the mountains, about half the land is under scrub of various kinds; and maquis, the classic Mediterranean scrub complex—with oleander, bay, evergreen oak, olive, and juniper—is particularly well developed in the Peloponnese. Evergreen trees and shrubs and herbaceous plants are found in the lowlands, with the flowers offering brilliant patterns in springtime. Pines, planes, and poplars line the rivers, the higher slopes, and the coastal plains. Oak, chestnut, and other deciduous trees are found in the north, giving way at higher altitudes to coniferous forests dominated by the Grecian fir, in which clearings are carpeted in spring and summer with irises, crocuses, and tulips. Forests and scrub are found at the highest levels; the black-pine forests covering Mount Olympus are especially noteworthy.

The forested zones, especially in the north, harbour such European animals as the wildcat, martin, brown bear, roe deer, and, more rarely, wolf, wild boar, and lynx. Animals of the Mediterranean regions include jackals, wild goats, and porcupines, all adapted to the lack of moisture and to the heat. Birds include pelicans, storks, and herons, while many varieties from farther north winter in Greece. Reptile and fish life is rich and varied.

Settlement patterns. In terms of human geography, Greece can be described as "classical Mediterranean" only in part, the other part being distinctly "Balkan." History rather than the physical environment accounts for fundamental paradoxes and contrasts in settlement pattern, social composition, and demographic trends that cannot be explained simply by reference to the difference between "Old Greece" and territories annexed in the early 20th century. For instance, although Greece is an "old country," relatively densely occupied in prehistoric times and well settled and much exploited in, and since, ancient times (as the large number of ancient monuments and important archaeological sites testifies), instability is as characteristic of Greece's settlement pattern as of its history. New villages, associated not only with Ottoman colonization but more recently (the first third of the 20th century) with agrarian reform, are juxtaposed with some of the most ancient towns of Mediterranean Europe (including Mycenea, Pilos, Thira, Árgos, Athens, Sparta, and Thebes). Traditionally, towns as well as villages have depended on the food potential of the surrounding land. This self-sufficiency, the *autarkeia* of the ancient city-states, survives in the remote villages, perforce traditional in their isolation, of mountainous Greece. Only Corinth and, above all, Athens were major trading centres in ancient times. The other major nuclei of trade were found where routeways (sea and land) coincided with cultivatable land. From the Byzantine period onward, fortification became an essential factor for monastic and secular settlement alike, emphasizing the importance of the mountain regions and of sites "perched" above lowland. As late as the 1960s, more than 40 percent of Greece's population lived in mountain regions. Intermittent periods of relative stability saw a return to the plains where the settlement pattern, dispersed or nucleated, often geometrically laid out, thus always seems to be "new."

Juxtaposition of new and old settlements

Greeks have preserved a strong sense of community. Village life remains a powerful influence, and village-square discussions reflect the cosmopolitan nature of the communities. This holds true despite the decline of the rural population in the late 20th century (still, more than one-third of Greece's total population is classified as rural). The same may be said about the small villages and towns

© John Sims—CLICK/Chicago



Greek Orthodox priests from the monastery of St. John the Theologian celebrating an outdoor Easter service on Pátmos, island of the Dodecanese group, Aegean Sea.

at the bottom of the urban hierarchy. At the other end of the urban scale, however, Greece's larger towns and cities have gained considerably in size and commercial importance since the 1970s. Athens, with a population of 750,000 increasing to about 3,000,000 for the entire metropolitan area (including the port of Piraeus), stands alone, but towns such as Thessaloníki, Patras, Vólos, Lárisa (Lárissa), and, on Crete, Iráklion are all fast-growing centres. Almost two-thirds of the population is now classified as urban, and another 10 percent as semiurban. Urbanization also is reaching out into the countryside, especially where excessive fragmentation of landholding (a consequence of agrarian reform) attracts urban-based financial and marketing entrepreneurs. Curiously, early Greek city planning, unlike Roman, has bequeathed little to the layout of modern urban centres.

THE PEOPLE

Linguistic, ethnic, and religious background. The inherent instability of the Balkan Peninsula—located as it is at the crossroads of invading Turks, migrating Slavs, and colonizing powers from western or central Europe (Venetians, Austro-Hungarians)—has bequeathed a bewildering amount of cultural confusion to Greece. Even in the south or on the islands, centuries of population migration and forced population exchanges continued well into the 20th century. Despite the long Ottoman administration (perhaps because of its failure to create a nation-state), all but a very small part of the population belong to the Church of Greece (Greek Orthodox church). This body appoints its own ecclesiastical hierarchy and is headed by a synod of 12 metropolitans under the presidency of the archbishop of Athens. The Greek church has links in dogma with the other Orthodox churches. Virtually all Cretans belong to a special branch of the Church of Greece, headed by the archbishop of Crete and directly responsible to the patriarchate of Constantinople.

The Greek Orthodox church

The Muslim minority, which constitutes most of the non-Orthodox group, is mainly Turkish and is concentrated in western Thrace and the Dodecanese. Roman and Greek Catholics, concentrated in Athens and the western islands formerly under Italian sway, account for the rest, except for a few thousand adherents of Protestant churches and of Judaism, the last group having been much reduced in numbers by the Nazi genocide of World War II.

In terms of ethnic composition, Greeks again make up all but a small part of the total, the remainder being composed of Macedonians, Turks, Albanians, Bulgarians, Armenians, and Gypsies. Except in Cyprus, southern Albania, and Turkey, there are no major enclaves of Greeks in nearby countries, although Greek expatriate communities play a distinctive role in western Europe, North and South America, and Australia.

Demography. The Greek population has never displayed high rates of growth, although—despite losses in a succession of wars and constant emigration as a result of poor economic conditions—it has usually shown a regular increase since the first census, in 1828. Most of its growth in the years since Greece gained its independence from the Turks in 1832 resulted from two factors—annexations of surrounding areas (the Ionian Islands; Thessaly and Árta; Epirus, Greek Macedonia, and Crete; Thrace; and the Dodecanese) and the influx of some 1,300,000 Greek refugees from Asia Minor in the 1920s. Emigration has continued to be a limiting factor: the years 1911–15 were an active period, and emigration became particularly heavy after World War II. The most common destinations of the emigrants have been the United States, Canada, Australia, and, somewhat later, Germany, Belgium, and Italy.

With a total population, according to the 1991 census, of 10,264,156, the two decades since the demographically stagnant 1950s and '60s have seen a remarkable revitalization in Greece. This is, however, almost wholly due to international population movements, not to an increase in natural growth rates, which remain low. Within the country, the contrast between regions losing population (two-thirds of the southern Peloponnese, all the Ionian isles except Corfu, the mountains of central, southwestern, and northeastern mainland Greece, and most of the islands of the eastern Aegean) and those rapidly gaining people (Attikí and other districts outside the major cities) holds a range of important social and political implications at all levels. (C.D.S.)

THE ECONOMY

Despite a rapid rate of growth in the post-World War II period, Greece's economy is one of the least developed in the European Union (EU). Natural resources are limited, industrialization has been achieved only partially, and there are chronic problems with the balance of payments. Shipping, tourism, and, decreasingly, migrant remittances are the mainstays of the economy. By the 1990s receipts from tourism amounted to one-quarter of the trade deficit.

Although the Greek economy has been traditionally based on free enterprise, many sectors of the economy have come under direct or, through the banks, indirect government control. This process of expanding state ownership of the economy has, historically, been associated as much with right-wing as with centre to left governments. Trade unions, which are fragmented and highly politicized, wield significant power only in the public sector. Measures were taken in the late 1980s and the early 1990s to diminish the degree of state control of economic activity. Following entry into the European Union, Greece has been a major beneficiary of subsidies for its generally inefficient agricultural sector and for infrastructural projects. Rates of productivity, however, remain low in both the agricultural and industrial sectors, and the development of the country's economy has lagged behind that of its EU partners. Unemployment, hitherto low, has grown as temporary migrants to other European countries have returned to Greece because of those countries' declining demand for immigrant labour. However, some sectors of the economy, notably shipping, have shown considerable dynamism.

Resources. Greece has few natural resources. Only in the case of nonferrous metals are there substantial deposits. Of these the most important is bauxite, reserves of which amount to more than 650 million metric tons.

Fossil fuels, with the exception of lignite of low calorific value, are in short supply. There are no deposits of bituminous coal, and oil production, based on the Prinos field near the island of Thasos, is very limited. The complex dispute between Greece and Turkey that developed in the 1970s over the delineation of the two countries' respective continental shelves—and hence the right to such minerals, in particular oil, as may exist under the Aegean seabed—shows no sign of being resolved.

Much of Greece's electrical power needs are supplied by lignite-fueled power stations and by hydroelectric power. Recently, attention has been given to the possibilities of solar and wind power.

Agriculture, forestry, and fishing. Greece's agricultural potential is hampered by poor soil, low rainfall, a system of landholding that has resulted in the proliferation of uneconomic smallholdings, and a general flight from the countryside to either the towns or overseas. About 30 percent of the land area is cultivable, the remainder consisting of scrub or forest. Only in the plains of Thessaly, Macedonia, and Thrace is cultivation possible on a reasonably large scale. Here corn (maize), wheat, barley, sugar beets, cotton, and tobacco are grown, Greece being a major EU producer of the last two items. Crops

Other crops grown in considerable quantities are olives (much of the annual crop being turned into olive oil), grapes, melons, peaches, tomatoes, and oranges, which are exported to other EU countries. Historically, Greek wine production, including the resin-flavoured retsina, has been primarily for domestic consumption, but efforts have been inititated to produce wines of higher quality for the world market.

Although inefficient, Greek agriculture has benefited substantially from EU subsidies, and there are many signs of growing rural prosperity. The importance of the agricultural sector to the economy, however, is diminishing.

Forests, mostly state-owned, cover approximately one-fifth of the land area, but they are subject to major forest fires. Forest products make no significant contribution to the economy.

Greece's huge coastline and numerous islands have given rise to a fishing industry. However, overfishing and failure to conserve fish stocks properly have lessened the contribution of fishing to the economy.

Industry. The industrial sector in Greece is weak. An established tradition exists only in the production of textiles, processed foods, and cement. (What is said to be the world's largest cement factory is located in Vólos.) In the past, private investment has been oriented much more toward real estate than toward industry, and concrete apartment blocks proliferate throughout the country. In the 1960s and '70s, taking advantage of an investment regime that privileged foreign capital, Greek shipowners invested significantly in sectors such as oil refining and shipbuilding. Shipping continues to be a key industrial sector, with the merchant fleet being one of the largest in the world, even if many of its ships are older than the world average. In the 1970s many ships that had hitherto registered under flags of convenience returned to the Greek flag. The fact that Greek ships, predominantly bulk carriers, are principally engaged in carrying cargoes between third countries renders the shipping industry vulnerable to downturns in international economic activity.

Tourism

Since the 1960s tourism has developed markedly, although Greece has not had much success in attracting high-spending tourists and is facing growing competition from Turkey. The number of tourists tripled between the early 1970s and the late 1980s. Most tourists come from other European countries. The emergence of a consumer society has created a seemingly insatiable demand for imported consumer goods, with negative consequences for the balance of trade. Road transport has improved immeasurably over the past 50 years, and there is a well-developed network of truck- and car-carrying ferries linking mainland Greece to the numerous islands and to Italy.

Finance. The central bank is the Bank of Greece. A significant number of the country's commercial banks are state-controlled. In the early 1990s banks controlled by the state held some 70 percent of total deposits. There is also a considerable degree of state control of the insurance sector.

In the early 1990s 118 public companies were quoted on the Athens stock exchange. For many Greeks, however, real estate, foreign currency, gold, and jewelry have proved a more attractive investment than stocks and shares. A pension and social insurance system of byzantine complexity is a major obstacle to economic modernization. The main social security fund, the Social Insurance Institute (IKA), is prone to recurrent crises in funding.

Trade. By the early 1990s some two-thirds of Greece's trade was with the other member countries of the European Union, the two main trading partners being Germany and Italy. Basic manufactures (*e.g.*, steel, aluminum, cement, and textiles), miscellaneous manufactured items (*e.g.*, clothing), and food (including livestock) each accounted for under one-quarter of exports; refined petroleum and petroleum-based products constituted a further 10 percent. Exports grew rapidly in the 1970s but slowed markedly in the '80s. Shipping and tourism contributed just over 10 percent to the gross domestic product (GDP) in the early 1990s, but there was a serious deficit in the balance of payments. This was offset by borrowing, limited foreign investment, and, to a decreasing extent, by emigrant remittances.

Transportation. Internal communications in Greece have, historically, been poor. Only during the post-World War II period have all the country's villages become accessible to wheeled traffic (and linked to the national electricity grid). There are no navigable rivers and only one canal, the Corinth Canal (completed in 1893), which divides the Peloponnese from mainland Greece. Although the canal significantly shortens the sea route from the Italian ports to Piraeus, the port of Athens, it has never fulfilled the economic expectations of its builders, because of its shallow draft and narrow width.

The Corinth Canal

Railway construction got under way in the 1880s, and, given the rugged terrain of the country, it involved some difficult feats of engineering. The total track is slightly under 1,600 miles in length, including the narrow-gauge railway network in the Peloponnese. The railway system is being modernized with the aid of EU funding. Trunk roads are inadequate by European standards, and Greece has one of the worst automobile accident records in Europe.

Public transport in the Athens metropolitan area is heavily dependent on an overcrowded and unreliable bus network. After many postponements, work on the much-needed Athens metro commenced in earnest in 1993. This will supplement the small suburban railroad network linking Athens' northern suburb of Kifisiá with the port of Piraeus.

The extensive internal bus-and-ferry network has been augmented since the 1960s by the development of a domestic flight network linking Athens with 25 domestic airports. The country's main airports are Ellinikón in suburban Athens (to be replaced in the late 1990s by an entirely new airport at Spáta in Attikí) and Macedonia, near Thessaloníki. Other international airports, which service the country's important tourist industry, are to be found on the islands of Crete (Iráklion), Corfu, Rhodes, Cos, and Lesbos and at Alexandroúpolis in Thrace and Andravida in the northwestern Peloponnese. The national carrier is Olympic Airways, which is 51-percent state-owned.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. *Constitutional framework.* The current constitution was introduced in 1975 following the collapse of the 1967–74 military dictatorship. The considerable powers it vouchsafed to the president were never invoked before they were reduced in the constitutional revision of 1986. Presidential powers are now largely ceremonial. The president is elected by the parliament (Vouli) and may hold office for two five-year terms.

The prime minister, who has extensive powers, must be able to command the confidence of the parliament. The latter consists of 300 deputies and is elected for a four-year term by direct, universal, and secret ballot. It has the power to revise the constitution, as happened in 1986. A distinctive feature of the electoral system is the practice of incumbent governments, of whatever political hue, amending the electoral law to suit their own political advantage. Voting is compulsory.

Parliament

The party system. Although the political system is in the process of modernization, many elements of traditional politics remain, notably the personalistic nature of the party system, with parties being heavily dependent on the charisma of their (frequently elderly) leaders and the importance of patronage at all levels.

There are three main political concentrations: the right, the centre, and the left. In the 1990s these were represented respectively by New Democracy (ND), the Panhellenic Socialist Movement (PASOK), and the Communist Party of Greece (KKE). New Democracy, founded by the veteran conservative politician Constantine Karamanlis, has progressively espoused "neoliberal," antistatist policies meant to limit the power of the state and to encourage private initiatives. PASOK, although it has substantially moderated the Third World liberationist rhetoric of its earlier years, retains a strong commitment to a radical foreign policy and an idiosyncratic form of socialism, which reflects the fact that only some 40 percent of the working population are wage or salary earners (the remaining 60 percent being self-employed). On the far left the KKE advocates a Soviet-style communism even after the demise of the Soviet Union. The broadly "Eurocommunist" Coalition of the Left and Progress has limited electoral appeal.

Local government. The country is divided into 13 geographic regions (9 mainland and 4 insular). These, in turn, are further subdivided into 51 departments (*nomoi*), each administered by a government-appointed prefect (*nomarkhis*). There is a government minister with special responsibility for Macedonia and Thrace.

The governmental system is highly centralized. The powers of local government are severely circumscribed by its inability to raise revenue.

Armed forces. The military has been a major arbiter of political life during the 20th century, but there has been no sign of political activity since 1974. Greece's expenditure on defense, at some 6 percent of GDP, is the highest in

the North Atlantic Treaty Organization (NATO) alliance. Conscription for men is universal, the period of compulsory active service lasting 19 to 23 months. Women have the right to volunteer for service.

Judicial system. The judicial system is essentially the Roman law system prevalent in continental Europe. The two highest courts are the Supreme Court (Areios Pagos), which deals with civil and criminal cases, and the Council of State (Symvoulion Epikrateias), which is responsible for disputes arising out of administration. A Court of State Auditors has jurisdiction in a number of financial matters. A Special Supreme Tribunal, whose members include the heads of the three courts mentioned above, deals with disputes arising out of the interpretation of the constitution and checks the validity of parliamentary elections and referenda.

Education. Education has long been prized in Greece both as an end in itself and as a means of upward social mobility. Wealthy Greeks of the diaspora have been major benefactors of schools and universities in their homeland. The educational system is somewhat rigid and heavily centralized, but the rate of literacy is high. Because of the inadequacies of state education, many children attend private *phrontistiria,* or institutions providing supplementary coaching outside normal school hours.

Competition for university places, which hold out the prospect of job security, is exceptionally severe. The oldest university-level institutions are the National Capodistrian University of Athens (1837), the National Technical University of Athens (1836), and the Aristotelian University of Thessaloniki (1925). This last institution has a tradition of innovation as compared with the more conservative University of Athens. From the 1960s to the '80s, a number of new universities were founded in Ioánnina, Patras, Thrace, Crete, Corfu, and the Aegean. However, they are often inadequately equipped and still do not offer a sufficient number of places to satisfy the demand for university-level education, forcing many Greek students to study abroad. Although there is a constitutional ban on private universities, a number of university-type institutions, some of dubious quality, have come into existence.

Health and welfare. Major strides have been made in the post-World War II period in eradicating diseases such as malaria, tuberculosis, typhoid, and dysentery. There are more doctors per person in Greece than in most of the other member countries of the European Union, and in the 1980s the PASOK government of Andreas Papandreou instituted a national health system. Many Greeks, however, where they can afford it, choose to travel abroad for medical care. Pension provision in Greece is a subject of extraordinary complexity. Some 80 percent of the working population are insured under the Social Insurance Institute and the Agricultural Insurance Organization (OGA; for farmers) programs.

Family law

During the 1980s important changes were introduced in Greek family law. Civil marriage was instituted in parallel with religious marriage, the dowry system was abolished (in theory), divorce was made easier, and the hitherto dominant position of the father in the family was restricted.

CULTURAL LIFE

The important sites of Greek antiquity that attracted European noblemen to the Greek lands in the 18th century, and which were such a potent influence on architectural styles in the West, continue to attract tourists from all over the world. Newly excavated sites such as the supposed tomb of Philip II of Macedon at Verghina and the Pompei-like remains at Thera are further indications of an astonishingly rich heritage from antiquity that has still not been fully explored. Over the past century there has been a greater awareness of the richness of the architectural and artistic heritage of the medieval empire of Byzantium.

The arts. Against the background of this extraordinary artistic heritage, Greece enjoys a thriving cultural life. It is in the field of literature that Greece has made its greatest contributions. Constantine Cavafy (1863–1933), who lived most of his life in Alexandria, Egypt, is frequently ranked among the great poets of the early 20th century. His poetry is suffused with an ironic nostalgia for Greece's past glories. Two Greek poets have won the Nobel Prize for Literature, George Seferis in 1963 and Odysseus Elytis in 1979. The novelist best known outside Greece is the Cretan Nikos Kazantzákis, whose *Zorba the Greek* was made into a popular film. A number of Greek composers have acquired an international reputation, including Nikos Skalkottas, Manos Hadjidakis, Mikis Theodorakis, and Iannis Xenakis, a French composer of Greek descent. Well-known painters in the post-World War II period include Ghika, Yannis Tsarouchis, and Photis Kontoglou, who drew his inspiration from the ascetic traditions of Byzantine art.

There is a lively theatrical tradition, in which political satire plays an important part. The traditional shadow puppet theatre, *Karaghiozis,* is now largely extinct, having been displaced by the ubiquitous television.

Cultural institutions. A thriving theatrical tradition is reflected in a myriad of theatres in the capital, whose repertoire ranges from Western classics to political satire. During the summer months huge audiences are attracted to performances of ancient Greek drama held in the theatre of Epidaurus, which dates from the 4th century BC and whose acoustics are extraordinary; the 2nd-century-AD Roman theatre of Herodes Atticus at the foot of the Acropolis in Athens also draws many visitors and is the location for concerts given in the framework of the annual Athens Festival held during the summer months. Live performance of orchestral music, limited in comparison with that of other European capitals, was given a major boost with the opening in 1991 of a newly built concert hall, the Megaro Mousikis (palace of music).

The Archaeological Service

Given the richness of the country's archaeological heritage and the emphasis in the country's official self-perception on continuity with the classical past, the Archaeological Service has assumed particular importance. Frequently working in cooperation with the various foreign archaeological institutes, it is responsible for excavating relics of the past and for running the country's museums. Public library provision is relatively limited, and there is no adequate national library. The country's most prestigious learned society is the Academy of Athens. A distinctive feature of intellectual life is the numerous societies devoted to the study of local and regional archaeology, history, and folklore, a development that reflects the strong regional loyalties of many Greeks.

Daily life. In the hot summers social life in Greece tends to be conducted outdoors. In small towns and villages the tradition of the *volta* continues, when much of the population strolls up and down the main street or, on the islands, the quayside at sundown. In summer and winter much leisure time is passed in the numerous cafés and coffee shops. These latter have traditionally been a male preserve, and it is not uncommon to find in a single village one coffee shop where the adherents of one political party congregate and another for supporters of the rival party. Television and other forms of video entertainment, however, threaten to undermine traditional leisure patterns.

Cuisine

The country's cuisine, particularly sweets such as *baklava* and *kataifi,* reflect the influence of the centuries of Turkish rule. The food in Thessaloníki, the capital of northern Greece, which was annexed to the Greek state only in 1912, reflects the Ottoman influence and is testimony to the massive influx of refugees from Asia Minor in the 1920s. These immigrants were often facetiously referred to by the native inhabitants as *yiaourtovaptismenoi* ("baptized in yogurt") on account of their fondness for yogurt in their appreciably superior cuisine. The traditional diet of the peasants was a healthy one based on vegetables, olives, olive oil, cheese, and bread, with meat being a luxury to be eaten only on special occasions. With growing affluence meat has come to assume a more important place in the country's diet, and the incidence of heart disease has risen accordingly.

Greek society is noted for its tight family structures and the low rate of crime. The extended family, and the obligation placed on family members to provide mutual support, is all-important. The centrality of the family has been little affected by the process of embourgeoisement that has been a characteristic feature of the development

of Greek society in the period since the end of World War II. Although the dowry system has officially been abolished, marriages still continue to be seen to a degree as economic alliances. The great majority of the country's businesses remain small, family-run enterprises. This is also true of ship-owning, the most dynamic sector of the economy. Tightly knit clans of ship-owning families dominate this industry. The family structure of industry acts as an impediment to modernization. The wheels of society continue to be lubricated by *mesa* (connections) and *rouspheti* (the reciprocal dispensation of favours).

The main holiday periods revolve around Easter and the Feast of Dormition (Assumption) of the Virgin in mid-August. Easter is the most important religious and family festival, with many people returning to their native villages for the traditional festivities, which include the vigil in church on Saturday evening, the lighting of the Holy Fire at midnight, and the roasting of whole lambs on the spit. August is the traditional holiday month. The national sport is football (soccer), and the fortunes of the principal teams are the focus of passionate loyalties. Hunting is another popular pastime.

Press and broadcasting. During the 1980s traditional newspaper proprietors were to an extent displaced by new entrepreneurs. Most newspapers became tabloids. The circulation of morning papers declined while that of evening papers increased. Leading newspapers include *Kathimerini* ("Daily"), *Eleftherotypia* ("Free Press"), and *Ethnos* ("Nation"). For the most part, newspapers tend to be unashamedly partisan in their political comments, with the laws of libel inspiring little fear in publishers. The government monopoly of television and radio broadcasting was broken in the 1980s. Private television and radio stations now exist in profusion. Like the press, broadcasting is unrestrained, particularly in its handling of political issues, although often at the expense of quality.

(R.R.M.C.)

History

Geographically, Greece forms, as previously noted, the southernmost extension of the Balkan Peninsula. It is a region dominated by mountain systems, and, although not particularly high, these cover some 80 percent of the surface area. The main formations are those of the Dinaric Alps, which push down from the western Balkan region in a southeasterly direction and which, in the Pindus Mountains, dominate western and central Greece. Extensions and spurs of these mountains form the salient features of southern Greece and the Peloponnese. The Balkan range lies north of Greece, extending eastward from the Morava River for about 340 miles (550 kilometres) as far as the Black Sea coast, but the Rhodope Mountains form an arc stretching from this range through Macedonia toward the plain of Thrace. The coastal and riverine plains are in consequence relatively limited in extent; moreover, they are differentiated by marked variations in climate, ranging from the Mediterranean type along the coast to the continental type inland, in the highlands. These plains reveal an accentuated settlement pattern consisting of a series of fragmented geopolitical entities, separated by ridges of the highlands, that fan out along river valleys toward the coastal areas. This structure played a significant role in shaping the history of preclassical and classical Greece and continued to do so in the medieval period; for, in spite of the administrative unity and relative effectiveness of the fiscal and military administration of the later Roman and Byzantine states, these still had to function in a geophysical context in which communications were particularly difficult. The southern Balkan Peninsula has no obvious geographic focal point. The main cities in the medieval period were Thessaloniki and Constantinople, yet these were peripheral to the peninsula and its fragmented landscape. The degree of Byzantine political control during the Middle Ages is clearly reflected in this. In the Rhodope Mountains, perhaps the most inaccessible of those mentioned, as well as in the Pindus Mountains, state authority, whether Byzantine or Ottoman, always remained a rather distant factor in the lives of the inhabitants. These were regions in which paganism and heresy could survive with little interference or control from a central government or church establishment.

Geophysical structure and land use

This geophysical structure also has affected land use. The highland regions are dominated by forest and woodland and the lower foothills by woodland, scrub, and rough pasture. Only the plains of Thessaly and Macedonia offer the possibility of extensive arable exploitation. The riverine plains and the coastal strips associated with them (such as the region around the Gulf of Argolis and, much more limited in extent, the Gulf of Corinth) present a similar but more restricted potential; they have been used for orchards, viticulture, and oleoculture. Inevitably, the pattern of settlement of larger urban centres as well as of rural communities is largely determined by these features.

Finally, the relationship between this landscape of mountains, gulfs, and valleys, on the one hand, and the sea, on the other, is fundamental to the cultural as well as to the political and military history of Greece. The sea surrounds Greece except along its northern bounds; and the extended coastline, including gulfs such as those of Corinth and Thessaloniki, which penetrate deep into the interior, has served as a means of communication with surrounding areas to the extent that even interior districts of the Balkans often share in the Mediterranean cultural world. The sea was also a source of danger: seaborne access from the west, from the south, or from the northeast via the Black Sea made Greece and the Peloponnese particularly vulnerable to invasion and dislocation.

GREECE DURING THE BYZANTINE PERIOD (*c.* AD 300–*c.* 1453)

Late Roman administration. At the beginning of the 4th century the regions comprising very approximately the modern state of Greece were divided among eight provinces: Rhodope, Macedonia, Epirus Nova, Epirus Vetus, Thessaly, Achaea, Crete, and the Islands (Insulae). (For the earlier history of Greece, see the article GREEK AND ROMAN CIVILIZATIONS, ANCIENT.) Of the eight provinces, all except Rhodope and the Islands were a part of the larger diocese of Moesia, which stretched up to the Danube River in the north; Rhodope belonged to the diocese of Thrace, while the Islands were classed as part of the diocese of Asiana, consisting, for the most part, of the westernmost provinces of Asia Minor. By the early years of the 5th century, administrative readjustments had divided the older diocese of Moesia into two sections, creating in the north the diocese of Dacia and in the south that of Macedonia, made up of the provinces of Macedonia I and II, with Epirus Novus and Epirus Vetus, Thessaly, Achaea, and Crete. Further changes during the middle of the 6th century resulted in the establishment of a military command known as the *quaestura exercitus*, a zone made up of the Islands and Caria, from the diocese of Asiana, together with the province of Moesia II on the Danube; it was designed as a means of providing for the armies based along the northern frontier in regions that were too impoverished or devastated to support them adequately.

Quaestura exercitus

In turn, these diocesan groups were parts of larger administrative units, the praetorian prefectures. Most of the Greek provinces were in the praetorian prefecture of Illyricum, except Rhodope, which, as a province of the diocese of Thrace, was in the prefecture of Oriens, as were the Islands. This pattern was radically altered by the developments of the 7th century.

Some of the ancient names for the regions of Greece disappeared from everyday use. However, many continued to be used in literary and administrative contexts, at least, especially in the administration of the church, or were revived by classicizing writers during the late Byzantine period. Thus, Aitolia, Akarnania, Achaia, Arkadia, and Lakedaimon were used in the 13th century and after. Similarly, in central Greece Boeotia, Euboea, and Thessaly all survived, in different contexts. Typical of their history is Euboea, which was so called until the 8th century, after which it was referred to variously as Chalkis or Euripus. After 1204 Western writers identified it as Negroponte, although the Byzantines also called it Euboea. The names Epirus and Macedonia seem never to have dropped out of



Greece in the early Byzantine period (7th and 8th centuries).

regular use. However, many new names also were coined during the Byzantine period; these tended to be geographic descriptions (such as Strymon or Boleron) used for both provincial and administrative divisions as well as to describe regions with a particular ethnic composition, for example, Vlachia in southern Thessaly.

The changing role of cities

The evolution of Byzantine institutions. As in other parts of the Roman world, the function of cities in the administrative structure of the state underwent a gradual evolution from the 3rd century on as the central government found it increasingly necessary to intervene in municipal affairs in order to assure itself of its revenues. This need for intervention developed when the vitality of cities became eroded as a result of a range of factors, most particularly the economic damage to urban infrastructures that accompanied the civil wars and barbarian inroads of the 3rd century. Further, the so-called "decline" of the curial order, that is to say, the weakening of the fiscal and economic independence of many towns by the tendency of members of urban elites to avoid municipal obligations, played a role. Finally, in the eastern parts of the empire, the transformation of Byzantium into Constantinople as a new imperial capital in 330 had important results for patterns of internal trade and commerce as well as for social relations between provincial elites and the state establishment. Nevertheless, the cities of the southern Balkans were able to survive the raids and devastation of both Goths and Huns in the 4th and 5th centuries, and there is no evidence that cities ceased to carry on their function, where it existed, as centres of market activity, local administration, and social life. Those cities that were artificial—*i.e.*, purely administrative creations of the Roman state—were the first to suffer, and, under difficult conditions, they generally disappeared. On the other hand, the state—through its regional and central military administration—appears also to have been involved in promoting different types of smaller defended urban centres, better adapted to these conditions.

In the 7th century a series of developments combined to further promote what was already in the process of becoming a radically different pattern of settlement and administration. In the first place, the results of the long-term developments already referred to, in terms of both social and economic stability, must be borne in mind. (Of particular significance is the fact that the municipal landowning elites transferred their attention—especially in respect of investment in the imperial system—away from their local cities to the imperial capital.) In the second place, the economic disruption brought about by the infiltration of large numbers of Slav settlers and immigrants was exacerbated by the devastation and insecurity caused by the wars between the empire and the Avars, a Turco-Mongol confederacy that was able, from the 560s to the 630s, to harness the resources of those peoples in the plain of Hungary and in the Balkans under its thrall to launch a series of extremely damaging and disruptive attacks on the empire. Although the disruptions of the earlier invasions of the Huns and Goths may have been as damaging in the short term, the Balkans were by this time suffering from the cumulative effects both of two centuries of constant insecurity and warfare and of endemic plague, which had struck the eastern empire in the 540s. It is clear from the imperial legislation of the late 5th to the 6th centuries dealing with the Danubian provinces and with Thrace, as well as from the archaeological record, that the final results of this economic disruption were to bring about the abandonment of the traditional pattern of urban economies. Even the construction of churches, which had experienced a certain efflorescence during the 6th century, virtually ceased in the 7th century; secular construction stopped; an extended process of ruralization of settlement and of economic life took place; and new populations penetrated far into the Peloponnese. The extent to which these new settlers replaced, or were able to live alongside, the indigenous population remains unclear.

The collapse of traditional administration

As a consequence of these changes, the traditional administration collapsed, chiefly because the government at Constantinople could control only the coastal plains and some river valleys. The cities that remained in imperial hands, such as Corinth or Thessaloníki, for example, were well-defended fortresses or had access to the sea, from where they could be supplied. Inland, tribal groupings and chieftaincies (*sklaviniai*) dominated many districts; and, while it appears that the empire claimed a political sovereignty over such regions, it was rarely able to make this effective or to extract regular revenues (although "tribute" was obtained at times). The *sklaviniai*, however, did not control the entire inland region; there were many areas in which the indigenous population and the traditional patterns of local social structure and political organization may have survived and where imperial authority may have been recognized. The evidence is too sparse to draw definite conclusions.

Evidence for the degree of Byzantine control over the

area is reflected dimly in the lists of signatories to ecclesiastical councils (especially those of 680 at Constantinople and of 787 at Nicaea) and in the various lists of bishoprics (*Notitiae episcopatuum*). From these, it is clear that ecclesiastical administration in the south, especially the Peloponnese, had suffered considerably. Many bishoprics were abandoned and had ceased to exist; at the council of 680 only 4 bishops from this region (Athens, Corinth, Lakedaimon, and Argos) and 12 from Macedonia attended; a (probably) 7th-century episcopal list (the *Pseudo-Epiphanios*) records the names of only 5 metropolitan bishops from Greece, mostly from the north.

Administratively, those districts that remained under Byzantine control were organized from the later 7th century onward into the military province, or theme (Greek: *thema*), of Hellas, under its general (*strategos*). The theme initially encompassed in all likelihood only the easternmost parts of central Greece but gradually came to include also parts of Thessaly and, possibly, of the Peloponnese, although in the latter case only the coastal regions were involved.

Hellas

The islands of the Aegean remained largely in imperial hands. In late antiquity they had been relatively heavily populated, the larger ones among them—especially Lemnos and Thasos in the north—being well-known sources of agricultural produce. Arab piracy and raiding from the later 7th century on altered this, causing many of the smaller islands to become deserted; however, the islands recovered during the 10th century.

Byzantine fortress-towns testify to the presence of sizable rural populations needing shelter from attack; the towns also served as refuges for people from the mainland fleeing Avar, Slav, or Bulgar raids. Administratively, they formed in the second half of the 7th century an element of the districts allotted to the naval theme (in the original sense of the term—*i.e.*, "army") of the Karabisianoi (Greek *karabos*, a light ship), which represented the rump of the *quaestura exercitus*. During the first half of the 8th century this command appears to have been subdivided to form the naval themes of the Kibyrrhaiotai (including parts of western Asia Minor and named after the district and town of Kibyrrha in southwestern Asia Minor), Samos or the Kolpos (Gulf) in the southern zone, and Aigaion Pelagos covering the northern districts. Further subdivisions took place in the 10th–12th centuries, so that commands of the Dodecanese or the Cyclades also appear in the sources.

It must be stressed that, even though the archaeological record clearly supports the conclusion that a dramatic collapse of traditional urban society and economic relations occurred, it also gives evidence of significant regional variations. The process of change was neither uniform across Greece, nor was it always as extreme in one area as in another. The degree of access to imperial resources and the ability to maintain regular contact with the imperial government were factors that gave some areas a very different appearance from others. It is important that this be borne in mind when considering the history of the various regions of Greece in the following period.

Byzantine recovery. The Byzantine recovery of effectively lost provinces began toward the end of the 8th century. The emperor Nikephoros I (802–811) is traditionally credited with a major role in this, although the process was certainly under way before his accession. The degree of Slavicization appears to have varied considerably. For example, it is clear that by the 10th century many districts of the Peloponnese were well Hellenized and thoroughly Christian once more. Yet outposts of Slavic language and culture—sometimes partly Hellenized, often relatively independent, in practice, of central imperial control—survived in the less-accessible regions until the 13th and 14th centuries and perhaps beyond. Traces of the preconquest social and political structures of the northern Peloponnese may be reflected in the story of the widow Danelis, a rich landowner whose wealth was almost proverbial in the later 9th century. She was a sponsor of the young Basil, later Basil I (867–886), and may have represented the last in a line of Christianized but semiautonomous Slavic princelings or chieftains who had dominated the region around Patras in Achaia.

Degrees of Slavicization

Different parts of Greece were reconquered at different times. Epirus in the northwest was gradually placed under Byzantine military administration, which was advancing inland from the coast during the first part of the 9th century. The themes of Cephalonia (Kefallinía) and Dyrrachion (Durazzo; modern Durrës) had been established by the 830s; that of Nikopolis appeared at the end of the 9th century. The theme of the Peloponnese emerged as a separate region in 812, although it was almost certainly created before this date; that of Thessaloníki had probably been established by about 812 (although this remains debated); those of Strymon and Boleron appeared likewise during the course of the 9th century.

Pattern of reconquest

Ecclesiastical organization once again reflects this process. A 9th-century list of bishoprics contains 10 Greek metropolitan sees, including those of Patras and Athens, compared with the 5 that appear in earlier records. And during the first half of the 8th century (in the context of the Iconoclastic Controversy, a religious controversy concerning the veneration of icons, or sacred images) the ecclesiastical provinces of the old prefecture of Illyricum, which had been subject to Rome, were withdrawn by the emperor Leo III (717–741) from papal authority and placed under Constantinople, thus permitting a unified program of re-Christianization of much of this region.

As Byzantine control became firmer, and as Byzantine military and political expansion northward accelerated during the 10th and early 11th centuries, older themes were subdivided, forming a mosaic of small administrative divisions. Thus the themes of Berroia, Drougoubiteia (clearly reflecting a Slavic tribal territory), Jericho (on the Adriatic coast between Dyrrachion and Nikopolis), and Edessa or Vodena (northwest of Thessaloníki) all appeared during the period from the later 9th to the 11th century.

Economy and society. Like other regions of the Byzantine Empire, Greece had suffered economically from the warfare of the 7th and 8th centuries. The rise of the khanate of the Bulgars, established south of the Danube after 681, whose rulers were able to exercise a hegemony over their politically fragmented Slavic neighbours, meant that warfare remained endemic and economic insecurity a factor of daily existence. Yet the restoration of Byzantine military and political power from the later 8th century onward and the growth of Byzantine cultural and religious influence throughout the Balkans during the 9th and 10th centuries created a context favourable to economic and demographic recovery throughout the empire, especially in the southern Balkan region. During the 11th and 12th centuries Greece experienced a powerful economic upswing, certainly more so than Anatolia. Cities such as Thessaloníki, Thebes, and Corinth became centres of flourishing local industries and of market exchange, rivaling the imperial capital in many respects. The silk industry that developed around Thebes was especially important. The evidence for greater wealth, and more especially greater disposable wealth in the hands of local elites, is found not only in documentary sources but also in a number of splendidly endowed churches, some still extant today. Many other towns, particularly those with a harbour or shelter for ships, became flourishing centres of trade and commerce and were sought-after locations for the trading posts of the Italian merchant republics after the 11th century.

Economic recovery

Results of the Fourth Crusade. The Fourth Crusade, called by Pope Innocent III to reconquer the Holy Land, was diverted to Constantinople. Following the crusaders' seizure and sack of the city in 1204, the European territories of the Byzantine Empire were divided up among the Western magnates. Whereas in Asia Minor Byzantine resistance was successful, so that two independent successor empires were established (those of Nicaea and Trebizond), most of Greece was quickly and effectively placed under Frankish (Western Christian) rule. The principality of Achaia (the Morea) and the duchy of the Archipelago were subject to the Latin emperor, the ruler of the Latin Empire (also referred to as Romania) set up in Constantinople in 1204 by the Latin (Western) Christians of the Fourth Crusade and claiming jurisdiction over the territories of the Byzantine state. A kingdom of Thessaloníki was es-



Greece *c.* 1215.

tablished, to whose ruler the lords of Athens and Thebes owed fealty; while the county of Cephalonia (which, along with the islands of Ithaca and Zákinthos had in fact already been under Italian rule since 1194, in the person of Matteo Orsini) was nominally subject to Venice, although it was in fact autonomous and after 1214 recognized the prince of Achaia as overlord. Finally, the lord of Euboia (Negroponte) was subject to the authority of both Thessaloníki and Venice. Byzantine control remained in the form of the despotate of Epirus in the northwest, in the area around Monemvasia in the eastern Peloponnese, and in the mountain fastness of the Taÿetos in Achaia and Arcadia. In 1261, however, the Nicaean forces were able to recover Constantinople and put an end to the Latin Empire. The recovery of some of the territory held by Frankish rulers followed, although Monemvasia actually fell, for a while, to a Frankish force in 1248: by the end of the 13th century parts of central Greece were once again in Byzantine hands, and the Byzantine despotate of Morea controlled much of the central and southeastern Peloponnese; but the principality of Achaia remained an important Frankish power to its north.

The history of Greece, as stated above, reflects very closely its geopolitical structure. This fact is particularly clear in the period following the Fourth Crusade, when the former Byzantine administrative divisions were organized into various petty states, each having its own local history and political evolution.

Despotate of Epirus. The so-called despotate of Epirus (ruled by a *despotēs*, or lord), which usually included Cephalonia, was established by Michael I Komnenos Doukas, who established effective control after 1204 over northwestern Greece and a considerable part of Thessaly. His brother and successor Theodore was able to retake Thessaloníki from the Latins in 1224, where he was crowned as emperor, thus challenging the emperors of Nicaea who claimed legitimate imperial rule. But in 1242 the Nicaean ruler John III Vatatzes compelled Theodore's son and successor John to abandon the title of emperor, and by 1246 Thessaloníki was under Nicaean rule. In 1259 much of Epirus came under Nicaean control, but this was lost by 1264; thereafter Epirus continued to be ruled by independent despots (*despotai*) until 1318. Its sheltered geographic position, between the spine of the Pindus range and the Adriatic, facilitated a degree of political separatism and independence from Constantinople until the Ottoman conquest. The Byzantine emperors, however, always insisted on their rights to confer the title of *despotēs*, and for much of the 14th and 15th centuries they regarded the rulers of Epirus as rebels.

From 1318 until 1337 Epirus was ruled by the Italian Orsini family, and, after a short Greek recovery, it was taken by the Serbs in 1348. Ioánnina and Arta were its main political centres. From 1366 to 1384 Ioánnina was ruled by Thomas Komnenos Palaeologus, also known as Preljubović, the son of the *caesar* Gregory Preljub, who had been Serbian governor of Thessaly under Stefan Uroš IV Dušan. He was able to assert Serbian control over northern Epirus and fought with the Albanian lords of Arta (Ghin Bua Spata and Peter Ljoša) in the south, eventually defeating them with Ottoman help. In 1382 his title of *despotēs* was confirmed by the Byzantine emperor at Constantinople. He was assassinated late in 1384, probably by members of the local nobility who objected to his rule. His widow, the Byzantine Maria Angelina Doukaina Palaiologina, married the Italian nobleman Esau Buondelmonti, who ruled as *despotēs* until about 1411. Thereafter the despotate came under the Italian house of Tocco, whose rulers were able to recover Arta from the Albanians. But in 1430 the Ottomans took Ioánnina, and Arta fell to them in 1449. Thenceforth Epirus was to be part of the Ottoman Empire. Cephalonia was taken in 1479, but Venice seized it in 1500.

Thessaly and surrounding regions. The political history of the other regions of Greece during this period is no less complex. Thessaly was ruled in its eastern parts by the Franks after 1204, while the western regions were disputed by the rulers of Epirus and Nicaea. About 1267 John I Doukas established himself as independent ruler, with the Byzantine title *sebastokrator*, at Neopatras, but in expanding his control eastward he came into conflict with Michael VIII, whose attacks he repelled with the assistance of the dukes of Athens and Charles I of Anjou. Venetian support, the result of a favourable trading relationship (Thessaly exported agricultural produce), helped maintain Thessalian independence until the arrival in 1309 of the Catalan Grand Company. This band of Spanish mercenaries, who originally had been hired by Andronicus II (1282–1328) to fight the Seljuks in Anatolia, had turned against imperial authority and established themselves in the Gallipoli peninsula. From there they moved into Greece through Thrace and Macedonia, which they plundered, and from 1318 onward they occupied the southern districts of Thessaly. The northern regions remained independent until 1332 (under the ruler Stephen Gabrielopoulos), after which they were taken by John II Orsini of Epirus. In 1335 Thessaly was retaken by the Byzantine Empire, and from 1348 it acknowledged the overlordship of the Serbian ruler Stefan IV. After his death (1355) the self-styled emperor Symeon Uroš, *despotēs* of Epirus and Akarnania, was able to seize control of both Epirus and Thessaly and rule independently following the death of Nikephoros II of Epirus in 1358/59. He was succeeded by his son John, who adopted the monastic life in 1373. The *caesar* Alexios Angelos Philanthropenos took control, governing as a vassal of the Byzantine emperor John V; but in 1393 the conquest of Thessaly by Ottoman forces put an end to its independence.

The Catalan Grand Company

Athens, Thebes, and Corinth. In the south, Greece was divided among a number of competing political units. After 1204 the dukes of Athens controlled much of central Greece, with their main base at Thebes. They had political interests to the north and in the Peloponnese. However, in 1311 the Catalan Grand Company established its power over the duchies of Athens and Thebes, turning out their Latin lords. Under the protection of the Aragonese king Frederick II of Sicily (three sons of whom became dukes of Athens), they dominated the region until the Navarrese Company (an army of mercenaries originally hired by Luis of Evreux, brother of Charles II of Navarre [1349–87], to help assert his claim over Albania and then temporarily in the service of the Knights Hospitalers, a military-monastic order) took Thebes in 1378 or 1379. This weakened Catalan power and opened the way for the Florentine Acciajuoli, lords of Corinth, to take Athens in 1388. The latter then ruled all three regions until their defeat at the hands of the Ottomans in the 1450s.

The Acciajuoli family

The Peloponnese. In the Peloponnese, the political rivalry between the Byzantines and the Frankish principality of Achaia dominated. The principality was at its most successful under its prince William II Villehardouin (1246–78), but in 1259 he had to cede a number of fortresses, including Mistra, Monemvasia, and Maina, to the Byzantines. Internecine squabbles weakened resistance to Byzantine pressure, especially from the 1370s onward, when Jacques de Baux hired the Navarrese Company to

fight for his claim to the principality. Upon his death in 1383 the Navarrese exercised effective political control over the Frankish territories under the commanders of the company. The last Navarrese prince, Pierre de St. Superan, joined the Ottomans in 1401 to raid Byzantine possessions in the southern Peloponnese; he died in 1402. He was succeeded by his widow, Maria Zaccaria, representative of an important Genoese merchant and naval family. She passed the title to her nephew Centurione II Zaccaria, who, however, lost much territory to the Byzantine despotate of the Morea. In 1430 he married his daughter to the Byzantine *despotēs* Thomas Palaeologus, handing over his remaining lands as her dowry. From this time on, the Byzantine despotate of the Morea effectively controlled most of the Peloponnese. However, the Ottoman presence and the fall of Constantinople to Sultan Mehmed II in 1453 effectively ended this final period of Byzantine rule. The Morea resisted Ottoman conquest until 1460, when it was finally incorporated into the Ottoman Empire (a year earlier than the empire of Trebizond, which fell in 1461). All of Greece was by this time under Ottoman authority, with the exception of some of the islands, which retained a tenuous independence under Venetian or Genoese protection.

Serbian and Ottoman advances. Byzantine power in the northern Greek regions was effectively destroyed by the expansion of the Serbian empire under Stefan Uroš IV Dušan (1331–55), the results of which included the loss of Epirus, Thessaly, and eastern Macedonia, as noted above. From the 1350s the Ottomans established themselves in Europe, taking the chief towns of Thrace in the 1360s and Thessaloniki in 1387. Apart from the despotate of the Morea, therefore, and certain of the Aegean isles, there remained in Greece no Byzantine imperial possessions by the beginning of the 15th century.

The Latin duchy of the Archipelago

The islands. A particularly complex picture is represented by the islands, which were a focus for the activities of the Seljuks and later the Ottomans, of the Venetians and Genoese, and of the Byzantines. Following the Fourth Crusade, much of the southern part of the Aegean came under Venetian authority; and, although Byzantine power was restored for a while in the late 13th century, Naxos remained the centre of the Latin duchy of the Archipelago, established in 1207 among the Cyclades by Marco Sanudo, a relative of the Venetian doge, with a body of freebooting merchants and nobles. Initially under the overlordship of the Latin emperor at Constantinople, the duchy later transferred its allegiance to Achaia (in 1261) and to Naples (in 1267), although Venice also claimed suzerainty. The Sanudo family was replaced in 1383 by the Lombard Crispi family, which retained its independence until 1566, when the duchy was conquered by the Ottomans (although ruled by an appointee of the sultan until 1579, when it was properly incorporated into the state).

The remaining islands were held at different times by the Venetians, the Genoese, the Knights Hospitalers, and, eventually, the Turks. Rhodes played a particular role in the history of the Hospitalers' opposition to the Ottomans. Until the early 13th century, the island had been in the hands of a succession of Italian adventurers, most of whom acknowledged the overlordship of the emperor at Nicaea. In 1308 the Hospitalers took control, having been based on Cyprus since 1291, the time of their expulsion from the Holy Land. Rhodes, however, fell finally in 1523, when the knights were permitted to remove to Malta. Of the northern Aegean islands, Lemnos remained Byzantine until 1453, before coming for a while under the rule of the Gattilusi of Lesbos (whose independence of the Ottomans finally ended in 1462). In 1460 it was awarded to Demetrios Palaeologus, formerly *despotēs* of the Morea, along with the island of Thasos (the latter having come under Ottoman domination in 1455). In 1479 it was occupied by Ottoman forces and formally incorporated into the Ottoman state. Other islands had equally checkered histories. Naxos and Chios fell only in 1566; complete Ottoman control was not achieved, indeed, until 1715, when Tenedos, which remained until that year under Venetian control, was taken.

The real exception to the Ottoman success in the Aegean, however, was Crete. Separately administered until the 820s, when it was seized by Spanish Arabs, it was conquered in 961 by the general and later Byzantine emperor Nicephoros II Phocas. After 1204 it was handed over to Boniface of Monferrat, who proceeded to sell it to Venice. Although oppressive and unpopular, Venetian rule witnessed the evolution of a flourishing Italo-Hellenic literary and political culture. It lasted until 1669, when, after a long siege of Candia and the creation and collapse of a range of temporary alliances between Venice and various western powers on the one hand and the Ottomans and their supporters on the other, the island passed into Ottoman hands. Crete

Economic and social developments. In spite of the political instability after 1204, Greece seems to have experienced relative prosperity in the later Byzantine period. Population expansion accompanied an increase in production as marginal lands were brought under cultivation, and trade with major and minor Italian mercantile centres flourished. Although hostility at the level of state politics was endemic, social relations between the ruling elites of Byzantine- and Latin-dominated areas were not mutually exclusive. Intermarriage was not uncommon, and a certain modus vivendi appears to have evolved. This contrasted with the attitude of the peasantry and ordinary population, whose perceptions were shaped by the Orthodox church, an identity as either Greek or Byzantine ("Roman"), and by hostility to the Western church and its ways. The Ottoman conquest was not seen as necessarily worse than Latin domination; in many cases, it was certainly welcomed as less oppressive.

Population and languages. One of the most vexing questions concerning the history of medieval Greece has been that of the extent to which the indigenous "Hellenic" population survived and brings with it the question whether this term can properly be used of anything other than a cultural (as opposed to ethnic or racial) identity. The archaeological data, certainly, can offer answers only in terms of cultural similarities and differences, so that the question, as it has been traditionally expressed, of a Hellenic ethnic survival, cannot be answered. The issue must be explored in the context of the influx of large numbers of Slavs during the later 6th–8th centuries as well as the migration across Greece of nomadic or semi-nomadic pastoral groups such as the Vlachs from the 10th or 11th century and the Albanians from the 13th century. Although the evidence of place-names suggests some lasting Slavic influence in parts of Greece, the evidence is qualified by the fact that the process of re-Hellenization that occurred from the later 8th century seems to have eradicated many traces of Slavic presence. Evidence of tribal names found in both the Peloponnese and northern Greece suggests that there were probably extensive Slavic-speaking populations in many districts; and from the 10th century to the 15th century Slavic occupants of various parts of the Peloponnese appear in the sources as brigands or as fiercely independent warriors. Whereas the Slavs of the south appear to have adopted Greek, those of Macedonia and Thessaly retained their original dialects, becoming only partially Hellenophone in certain districts.

The Albanians. The origins of the Albanians (*Albanoi/Arvanitai* in Greek) remain uncertain. They appear to be the descendants of the Illyrian populations who withdrew into the highlands of the central Dinaric chain. Their name may originate from the valley of the Arbanon (along the Shkumbi River) in the theme of Dyrrachion (Durrës/Durazzo), in which they were first noted by outside commentators. Their language probably evolved from ancient Illyrian (formerly classed with the Hellenic group of Indo-European languages but now generally recognized as an independent member of the latter family), but it is heavily influenced by Greek, Slavic, and Turkish, as well as medieval Italian. For reasons not yet fully understood, the Albanians began in the 14th century to advance into the western coastal plain, where they served both Byzantine and Serbian overlords as well as ruling independently under various warlords and chiefly families; they were also present in considerable numbers in Thessaly, Boeotia, Attica, and the Peloponnese, serving as soldiers and as

farmers, colonizing deserted lands. Albanians arrived in large numbers in the Peloponnese during the reign of the *despotēs* Manuel Kantakouzenos, who brought them there to serve as soldiers and to resettle depopulated regions. The impact of their presence on the region's existing ethnic and linguistic structure remains debated.

The Vlachs. In central and southern Thessaly, the Vlachs played an important role. They have generally been identified with the indigenous, pre-Slav populations of Dacian and Thracian origin, many of whom migrated into the less-accessible mountainous areas of Greece and the northern Balkan region because of the Germanic and Avar-Slav invasions and immigration of the 5th–7th centuries. The Vlachs maintained a transhumant, pastoral economy in those areas. Their language belongs to the so-called Macedo-Romanian group and is closely related to that known from the 13th century on as Romanian (Daco-Romanian); it is essentially rooted in the late Latin, heavily influenced by the Slavic dialects with which the Daco-Thracian populations were in regular contact. By the 11th century the Vlachs are described as communities of shepherds who moved with their flocks between their winter pastures in Thessaly and summer pastures of the Gramoz Mountain and Pindus range; they are found in Byzantine armies and are mentioned in many documents dealing with landholdings in northern Greece, where—as is often the case in relations between settled and nomadic populations—they were regarded as troublemakers and thieves. Byzantines were often imprecise in their use of ethnic names; the Vlachs seem frequently to have been confused with the Bulgarians, through whose territory they also wandered on their seasonal routes and pasturage. A major modern debate about the role of the Vlachs in the establishment of the Second Bulgarian empire after 1185 continues, strongly marked by nationalist sentiment.

The language of the Vlachs

Emerging Greek identity. As the Byzantine Empire declined, the predominant role of Greek culture, literature, and language became ever more apparent. For Christians of the early and middle Byzantine worlds, the terms Hellene and Hellenic generally (although not exclusively, since in certain literary contexts a classicizing style permitted a somewhat different usage) had a pejorative connotation, signifying pagan and non-Christian rather than "Greek." From the 12th century, however, in the context of increasing conflict with Western European culture on the one hand and the encroaching Turkish powers on the other, this situation changed. With the territorial reduction of the empire to strictly Greek-speaking areas, the multiethnic tradition gave way to a more self-consciously "national" Greek consciousness, and a greater interest in "Hellenic" culture in a positive sense evolved. Byzantines began to refer to themselves not just as Rhomaioi, the traditional term, but as Hellenes. Notions of nationhood developed, and among learned circles a deep interest in the classical past was cultivated. While there was a powerful secularist tradition in this, culminating in the ideas of the neoplatonic Byzantine philosopher George Gemistus Plethon (who argued for the implementation of the political-philosophical system outlined in Plato's *Republic*), it was the combination of popular Orthodoxy (and strongly anti-Western ecclesiastical sentiment) with a specifically Greek identity that shaped the Byzantines' notions of themselves in the twilight years of the empire. With the political extinction of the empire, it was the Greek Orthodox church and the Greek-language community, in both "Greece" and Asia Minor that continued to cultivate this identity as well as the ideology of a Byzantine imperial heritage rooted in both the Roman and the classical Greek past. (J.F.H.)

GREECE UNDER OTTOMAN RULE

Constantinople fell to the Ottoman Turks on May 29, 1453. The Byzantine emperor, Constantine XI Palaeologus, was last seen fighting alongside his troops on the battlements; his death gave rise to the widely disseminated legend that the emperor had turned to marble but that one day he would return to liberate his people. By 1453 the Byzantine Empire was but a pathetic shadow of its former glories. The fall of this symbolic bastion of Christendom in the struggle against Islām may have sent shock waves through Western Christendom, but the conquest was accepted with resignation by many of the inhabitants of the city; as they saw it, their plight was a consequence of the sinfulness of the Byzantine Empire. Moreover, for many people Ottoman rule, and the maintenance of the integrity of the Orthodox faith, was preferable to accepting the pretensions of the papacy, the price Western Christendom had sought to exact in return for military assistance to ward off the Turkish threat. What was more, it was widely believed that the end of the world would come in 1492.

The millet system. With the conquest of the territories that had constituted the Byzantine Empire, the Ottoman sultans were faced with the problem of governing large non-Muslim populations. Christians and Jews, as "People of the Book," were afforded a considerable degree of toleration. Indeed, it was to the Ottoman Empire rather than Christian Europe that many Spanish Jews migrated following their expulsion from Spain in 1492. The Ottomans confronted the problem of the governance of these large heterodox and polyglot populations by establishing millets. These were organized on the basis of religious confession rather than ethnic origin, of which, in any case, in the early centuries of Ottoman rule there was little consciousness. The ruling millet within the empire was made up of the Muslims. Next in importance was the Orthodox Christian *millet-i Rūm,* or "Greek" millet, as it was known. There was also an Armenian, a Jewish, a Roman Catholic, and even, in the 19th century, a Protestant millet. Although its head, the ecumenical patriarch, was invariably of Greek origin, the term "Greek" millet was something of a misnomer, for it included, besides the Greeks, Romanians, Bulgarians, Serbs, Albanians, Vlachs, and substantial Arab populations. With the rise of nationalism in the 18th and 19th centuries the non-Greek members of the "Greek" millet came increasingly to chafe at the Greek stranglehold on the higher reaches of the hierarchy of the Orthodox church, through which the millet was administered.

The ecumenical patriarch

The powers of the ecumenical patriarch were indeed extensive, although there is uncertainty as to the precise nature of the privileges granted by Sultan Mehmed II to the man whom he elevated to the highest office in the church. This was Gennadios II Scholarios, a known opponent of those who, in the last years of the Byzantine Empire, had advocated union with the Western church. Patriarchal authority was considerable and extended to civil as well as to strictly religious matters. In many respects, indeed, it was greater than that enjoyed by the patriarchs in Byzantine times. The privilege of a considerable degree of autonomy in directing the affairs of the millet carried with it the responsibility of ensuring that its members were unshaken in their loyalty to the Ottoman Porte, or government. If disloyalty manifested itself, then retribution was swift and harsh, as occurred at the time of the outbreak of the War of Greek Independence in 1821 when the patriarch Grigorios V was executed in reprisal, despite the fact that he had vigorously condemned the insurgents. In the West this was seen as an act of mindless barbarity. In the eyes of the Ottomans, however, Grigorios had signally failed to carry out his fundamental obligation, that of ensuring that the Orthodox flock remained loyal subjects of the sultan.

Disadvantages for non-Muslims. In keeping with Islāmic tradition, members of the Greek millet enjoyed a considerable degree of autonomy in conducting their religious affairs. They were nonetheless at a disadvantage in a number of ways in comparison with members of the ruling Muslim millet. A Christian was not allowed to bear arms and was disbarred from military service (although this latter disability was in many ways a privilege), in lieu of which he had to pay a special tax, the *haraḍ.* In a court of law, a Muslim's word was always accepted over that of a Christian, although disputes between Christians were generally settled in courts under the control of the millet. A Christian could not marry a Muslim woman, and there was a strict prohibition against apostasy from Islām. Indeed, Christians who had embraced Islām and then reverted to Christianity, were, until well into the 19th century, invariably punished by death. These "neomartyrs," however, helped sustain the faith of the Orthodox populations during the centuries of Ottoman rule.

The janissary levy

The most serious disability to which Christians were subject, until the practice died out toward the end of the 17th century, was the janissary levy (*paidomazoma*). Christian families in the Balkans were required, at irregular intervals, to deliver to the Ottoman authorities a given proportion of their most intelligent and handsome male children to serve, after being forcibly converted to Islām, as elite troops or civil servants. Inevitably the levy was much feared, but those conscripted frequently rose to high office and were sometimes able to help relatives or their native villages. Indeed, there is evidence of Muslim families seeking to pass their children off as Christian in the hope that they would be included in the levy and would thus be able to better their prospects. Under such pressures there were numerous instances of Christian conversion to Islām on both an individual and a mass basis, such conversions being particularly prevalent in the 17th century. Not infrequently, however, the conversions were only nominal, and these crypto-Christians secretly practiced the rituals of their former faith.

Resistance to Ottoman rule. During much of the four centuries of the "Tourkokratia," as the period of Ottoman rule in Greece is known, there was little hope that the Greeks would be able to free themselves by their own efforts, although there were sporadic revolts, such as occurred on the mainland and in the islands of the Aegean in the aftermath of the defeat inflicted on the Ottoman navy in 1571 by Don John of Austria; the short-lived revolt launched by Dionysius Skylosophos in Epirus in 1611; and the abortive uprising in the Peloponnese in 1770 at the time of the Russo-Turkish war of 1768–74. These uprisings had little chance of success, but throughout the centuries of the Tourkokratia there was a kind of armed resistance to the Turks in the form of the klephts (literally robbers). In their banditry the klephts did not distinguish between Greek and Turk, but their attacks on such manifest symbols of Ottoman authority as tax collectors led to their being seen as acting on behalf of the Greeks against Ottoman oppressors. Certainly they are viewed in this light in the corpus of klephtic ballads that emerged, extolling the bravery and military prowess of the klephts and their heroic resistance to the Ottomans.

In an effort to counter the depredations of the klephts and to control the mountain passes that were their favoured area of operation, the Ottomans established a militia of *armatoloi*. Like the klephts, these were Christians, and the distinction between klepht and *armatolos* tended to be a narrow one. Moreover, the existence of such armed formations meant that, when the Greek revolt broke out in 1821, there was an invaluable reserve of military talent.

The Leo oracles

Belief in divine intervention. Greek aspirations for freedom were largely sustained by a corpus of prophetic and messianic beliefs that foretold the eventual overthrow of the Turkish yoke as the result of divine rather than human intervention. Such were the oracles attributed to the Byzantine emperor Leo VI the Wise (886–912), which foretold the liberation of Constantinople 320 years after its fall—*i.e.*, in 1773. Particular credence was placed in this prophecy, for its fulfillment coincided with the great Russo-Turkish war of 1768–74, one of the periodic confrontations between the two great regional powers. The Russians were the only Orthodox power not under foreign domination, and they were widely identified with the legendary "xanthon genos," a fair-haired race of future liberators from the north. The Russians were seen as forming part of a commonwealth linking the various parts of the Orthodox Christian world, with its common centres of pilgrimage such as the monastic republic of Mount Athos (forming one of the three fingers of the Chalcidice Peninsula) and Jerusalem.

The role of the Orthodox church. The Orthodox church was the only institution to which the Greeks could look as a focus. Through the use of Greek in the liturgy and through its modest educational efforts, the church helped to a degree to keep alive a sense of Greek identity, but it could not prevent Turkish (which was written with Greek characters) from becoming the vernacular of a substantial proportion of the Greek population of Asia Minor and, indeed, of the Ottoman capital itself.

Competition for high office

The Orthodox church, however, fell victim to the institutionalized corruption of the Ottoman system of government. The combining of civil with religious power in the hands of the ecumenical patriarchate and the upper reaches of the hierarchy prompted furious competition for high office. This was encouraged by the Ottomans, for it was soon the norm for a huge *peshkesh*, or bribe, to be paid to the grand vizier, the sultan's chief minister, on each occasion that a new patriarch was installed. Thus, despite the fact that, in theory, a patriarch was elected for life, there was a high turnover in office. Some even held the office more than once. Grigorios was executed by the Ottomans in 1821 during his third patriarchate, while during the second half of the 17th century Dionysius IV Mouselimis was elected patriarch no fewer than five times. It was this kind of behaviour that prompted an 18th-century Armenian chronicler to taunt the Greeks that they changed their patriarch more frequently than they changed their shirt.

Bribes had to be paid to secure office at all levels, and these could be recouped only through the imposts placed on the Orthodox faithful as a whole. The clergy's reputation for rapacity led to the growth of anticlericalism at a popular level and, in particular, among the small nationalist intelligentsia that emerged in the course of the 18th century. The anonymous author of that fiery nationalist polemic the "Ellinikhi Nomarkhia" ("Hellenic Nomarchy"; 1806) was a bitter critic of the sloth and self-indulgence of the higher clergy, while Adamántios Koraïs (1748–1833), the intellectual mentor of the national revival, though careful to steer between what he termed the Scylla of superstition and the Charybdis of atheism, was highly censorious of the obscurantism of the clergy. What particularly incensed Koraïs and his ilk was the willingness with which the Orthodox hierarchy identified its interests with those of the Ottoman authorities. However, the views of men such as Anthimos, the patriarch of Jerusalem, who argued in 1798 that the Ottoman Empire was part of the divine dispensation granted by God to protect Orthodoxy from the taint of Roman Catholicism and of Western secularism and irreligion, were by no means unusual.

TRANSFORMATION TOWARD EMANCIPATION

Signs of Ottoman decline. During the 16th and 17th centuries the main preoccupation of the Greeks was with mere survival. In the course of the 18th century, however, a number of changes occurred both in the international situation and in Greek society itself that cumulatively gave rise to hopes that the Greeks might themselves launch a revolt against Ottoman authority with some promise of success. By the end of the 17th century the protracted process of Ottoman decline was clearly under way. The failure of the great siege of Vienna in 1683 signaled the beginning of a slow process of retreat in the European provinces of the empire. The military triumphs of earlier centuries gave way to pressure on the empire from the Austrians, the Russians, and the Persians. The Russian threat reached its apogee in the 1768–74 war with Turkey. The Russians were subsequently to claim the right to exercise a protectorate over all the Orthodox Christians of the Ottoman Empire on the basis of an optimistic reading of the terms of the peace settlement with the Ottoman Empire by the Treaty of Küçük Kaynarca.

As the empire lost territory as it was forced onto the defensive, so the control of the Ottoman Porte over its still enormous provinces weakened. In both European and Asiatic Turkey, provincial warlords usurped the authority of the sultan, and the example of successful defiance of the Porte afforded by powerful satraps such as Ali Paşa Tepelenë, the Muslim Albanian who ruled over a large swathe of mainland Greece, gave encouragement to Greek nationalists, for it demonstrated that the empire was no longer the invincible monolith it once had been.

The Phanariotes. Of critical importance to the ultimate success of the national movement was the profound transformation that Greek society was to undergo during the course of the 18th century. Significant among these developments was the rise to power and influence of the Phanariotes, a small caste of Greek (and Hellenized

Romanian and Albanian) families who took their collective name from the Phanar, or Lighthouse, quarter of Constantinople, the home of the ecumenical patriarchate. The roots of their ascendancy can be traced to the need of the Ottomans for skilled negotiators as the power of their empire declined. No longer in a position to dictate peace terms to their vanquished enemies, they now had to rely on diplomats skilled in negotiation who might mitigate the consequences of military defeat, and these were drawn from the Phanariotes. Between 1699, when the peace treaty with the Habsburg monarchy was signed at Carlowitz, and 1821, the year of the outbreak of the War of Greek Independence, Phanariote grandees monopolized the post of chief interpreter to the Porte. This was a more important post than it appeared, for its holder bore considerable responsibility for the conduct of foreign policy. Similarly, Phanariotes were invariably interpreters to the *kapudan pasha,* the admiral of the Ottoman fleet. Again their powers were wider than the title suggests, for these Phanariotes in effect acted as governors of the islands of the Aegean archipelago, whose Greek inhabitants were a principal source of the sailors manning the Ottoman fleet.

Hospodar

The most important posts held by Phanariotes were those of *hospodar,* or prince, of the Danubian principalities of Moldavia and Wallachia. Phanariotes ruled these potentially rich provinces as the viceroys of the sultans, and their sumptuous courts in Jassy (now Iaşi, Rom.) and Bucharest aped on a lesser scale the splendour of the imperial court in Constantinople. Just as there was furious, and corrupt, jockeying for high office in the Orthodox church, so the appointment of the *hospodar*s was accompanied by intrigue and corruption. Just as there was a high turnover in office of the ecumenical patriarch, so the average tenure in office of a Phanariote *hospodar* was less than three years. Because they needed to recoup their expenditures on bribes, *hospodar*s acquired a not wholly justified reputation for greed and oppression. Some *hospodar*s displayed an enlightened interest in legal and land reform. Most acted as patrons of Greek culture, education, and printing. The princely academies attracted teachers and pupils from throughout the Orthodox commonwealth, and there was some contact with intellectual trends in Habsburg central Europe. For the most part the Phanariotes were too closely wedded to the Ottoman system of government, of which they were major beneficiaries, to play a significant part in the emergence of the Greek national movement. Nonetheless, however much their interests coincided with the maintenance of the Ottoman status quo, they provided a pool of individuals with experience in diplomacy and politics when armed struggle erupted in 1821.

The mercantile middle class. The single most important development in the Greek world during the 18th century was the emergence of an entrepreneurial, prosperous, and far-flung mercantile middle class, which played a major role in the economic life of the Ottoman Empire but which was also active outside its bounds. Discouraged from investing their capital within the empire by the arbitrariness and rapacity of the state, these Greek merchants played an active role in developing commerce in Hungary and Transylvania, newly acquired by the Habsburg monarchy, and also in southern Russia, where they were encouraged to settle by the empress Catherine the Great after Russia's borders had reached the Black Sea. Greek became the basic language of Balkan commerce as these merchants challenged the existing hold of British, French, and Dutch merchants on the import-export trade of the empire, importing Western manufactured goods and colonial produce and exporting raw materials. Greek merchant communities, or *paroíkies,* each with their own church, were established through much of central Europe, the Mediterranean littoral, and southern Russia, and even as far afield as India.

The merchant marine

Paralleling this development, a substantial merchant marine, based on the three "nautical" islands of Hydra, Spetsai, and Psará, came into existence. This merchant marine prospered from running the continental blockade imposed by Great Britain during the period of the French revolutionary and Napoleonic wars. The existence of a reservoir of trained sailors was to prove of inestimable advantage once the war of independence had broken out, when Greek fire ships became a formidable weapon against the cumbersome ships of the line of the Ottoman fleet.

The emergence of a mercantile middle class had a number of important consequences. Greeks were brought into contact with the ordered societies of western Europe, in which the state gave encouragement to commerce. They compared this state of affairs with that prevailing in the Ottoman Empire, where the absence of the rule of law and general arbitrariness militated against the generation of capital and still more its retention. Most of the merchants were, like the Phanariotes, too wedded to the status quo to give active encouragement to the national movement and thus potentially threaten their newfound prosperity. However, indirectly at least, they made a major contribution to the emerging national movement. For it was their wealth that provided the material basis for the intellectual revival that was such a significant feature of the last three decades of the 18th century and first two of the 19th. Impelled by the sense of local patriotism that had always been strong in the Greek world, they endowed schools and libraries. It was no accident that the three most important schools-cum-colleges in the Greek world on the eve of the war of independence were situated in Smyrna, Khíos, and Ayvalik (on the coast of Asia Minor opposite the island of Lesbós), all three major centres of Greek commerce.

The intellectual revival. A significant number of schoolteachers studied, with the financial backing of their merchant benefactors, in the universities of western Europe, particularly those of Italy and the German states. There they came under the influence of the ideas of the European Enlightenment and encountered the intoxicating nationalist doctrines emanating from the French Revolution. Above all, they became aware of the reverence in which the language and culture of ancient Greece were held throughout Europe. This realization kindled in them a consciousness of their own past, a recognition of being the heirs to this same civilization and of speaking a language that had changed remarkably little in the two and a half millennia since the time of Pericles. During the 50 years or so before 1821 a veritable flood of books on the language, literature, and history of the ancient Greek world was published (albeit for the most part outside the Greek domains) for a Greek readership.

Adamántios Koraïs

A leading role in the rediscovery of the past was played by Adamántios Koraïs. A native of Smyrna, where he was born in 1748, Koraïs sought, unsuccessfully, to establish himself as a merchant in Amsterdam. After studying medicine at the University of Montpellier, he moved to Paris in 1788, where he soon experienced the French Revolution firsthand. The main interest of his life, however, was classical philology, of which he became one of the foremost scholars in the Europe of his day. He devoted his years in Paris (until his death in 1833) to the study of this subject as well as to inspiring in his compatriots an appreciation of their classical ancestry. With the help of a family of rich merchants of Jannina (Ioánnina) he published a whole series of editions of classical authors, which he prefaced with exhortations to his compatriots to cast off their Byzantine ignorance by reviving the glories of the ancient world and by imitating the French, the people of modern Europe who, in his estimation, most resembled his classical forebears. His panacea for the degraded condition of the Greeks was education; it would enable them to free themselves from the double yoke of the Ottoman Turks and the Orthodox church.

One consequence of what could be described as an obsession with antiquity on the part of the small nationalist intelligentsia was the practice of naming children (and ships) after the worthies of ancient Greece, a custom dating from the first decade of the 19th century. Another was the vigorous debate that got under way, and which has lasted until the present day, as to the form of the language that was appropriate to a regenerated Greece. Some advocated using the spoken language, the demotic, as the language of educated discourse. Others favoured the Katharevousa, or purified Greek, which would render it more akin to the supposed purity of Attic Greek. Still others, such as Koraïs, advocated a middle path.

The intellectual revival of the half century or so before 1821 was real enough. However, much of it took place in the Greek communities of the diaspora, and the nationalist enthusiasms of the intelligentsia left the great mass of the peasantry, most of whom were illiterate, largely unmoved. Moreover, the elites of preindependent Greek society, the higher clerics, the wealthy merchants, the Phanariotes, and the *kodjabashis*, the wealthy provincial notables, whose lifestyle sometimes led to their being derisively referred to as "Christian Turks," were for the most part supporters of the status quo under the Ottomans. Whatever the faith pinned by Korais on education, cultural revival by itself was not going to remove the incubus of the Turks.

FROM INSURGENCE TO INDEPENDENCE

Rigas Velestinlis. Toward the end of the 18th century Rigas Velestinlis (also known as Rigas Pheraios), a Hellenized Vlach from Thessaly, began not only to dream of, but actively to plan for, an armed revolt against the Turks. Rigas, who had served a number of Phanariote *hospodars* in the Danubian principalities, spent part of the 1790s in Vienna. There he had come under the influence of the French Revolution, as is manifest in a number of revolutionary tracts he had printed, intending to distribute them in an effort to stimulate a Pan-Balkan uprising against the Ottomans. These tracts included a "Declaration of the Rights of Man" and a "New Political Constitution of the Inhabitants of Rumeli, Asia Minor, the Islands of the Aegean, and the principalities of Moldavia and Wallachia." The latter proposed the establishment of what, in essence, would have been a revived Byzantine Empire, but an empire in which monarchical institutions would have been replaced by republican institutions on the French model. Rigas' insistence on the cultural predominance of the Greeks, however, and on the use of the Greek language, meant that his schemes had little potential interest for the other peoples of the Balkan Peninsula. In any case, Rigas' ambitious schemes came to naught. Before he had even set foot on Ottoman soil, he was betrayed by a fellow Greek to the Habsburg authorities, who promptly handed him and a small group of coconspirators over to the Ottoman authorities; he was strangled by them in Belgrade in the summer of 1798. At one level Rigas' conspiracy had thus been a miserable failure, but his almost single-handed crusade served as an inspiration to subsequent generations of Greek nationalists.

Western encroachments. The arrest of Rigas thoroughly alarmed both the Ottoman authorities and the hierarchy of the Orthodox church, for it almost coincided with the occupation of the Ionian Islands in 1797 by the forces of revolutionary France and with Bonaparte's invasion of Egypt in 1798. These developments occasioned panic in Constantinople, for they seemed to indicate that the seditious and atheistic doctrines of the French Revolution had arrived at the very borders of the empire. The brief period of French rule in the Ionian Islands, attended as it was by the rhetoric of revolutionary liberation, soon gave way to a short-lived Russo-Turkish condominium, a further period of French rule, and finally, after 1815, the establishment of a British protectorate. Although governed like a colony, the Ionian Islands under British rule in theory constituted an independent state and thus afforded an example of free Greek soil, adjacent to but not under the control of the Ottoman Empire.

Philiki Etaireía. The example of Rigas Velestinlis was very much in the minds of the three young Greeks, lowly members of the Greek mercantile diaspora, who in 1814 in Odessa in southern Russia, the centre of a thriving Greek community, founded the Philiki Etaireía, or "Friendly Society," with the specific aim of laying the foundations for a coordinated, armed uprising against the Turks. The three founders, Emmanuil Xanthos, Nicholas Skouphas, and Athanasios Tsakalov, had little vision of the shape of the independent Greece for which they aimed beyond the liberation of the motherland.

Initiation rituals

The initiation rituals of the Philiki Etaireía were strongly influenced by those of the Freemasons. There were four categories of membership, ranging from the lowly *vlamis* (brother) to the *poimin* (shepherd). Those who betrayed the conspiracy were ruthlessly dispatched. Initially the society's attempt to recruit members throughout the Greek world met with little success, but from 1818 onward it made some headway, finding the communities of the diaspora an important source of recruits. From the outset the leadership of the society, well aware that the overwhelming majority of the Greek people looked upon their fellow Orthodox, the Russians, as their most likely liberators, put it about, quite misleadingly, that the conspiracy enjoyed the support of the Russian authorities.

In this connection, two attempts were made to recruit as leader of the conspiracy Count Ioánnis Kapodístrias, a Greek from Corfu, who, since 1816, had served as joint foreign minister to Tsar Alexander I of Russia and who was well versed in the ways of European diplomacy. The conservative Kapodístrias, however, was dismissive of the plot and urged the Greeks to bide their time until there was yet another of the regular wars between the Russian and Ottoman empires, when they might hope to achieve the kind of quasi-autonomy enjoyed by Serbia since 1813. Although he could see no future in the plans of the members of the Philiki Etaireía, Kapodístrias did not betray the secret of the conspiracy. It was to another Greek in the Russian service, Prince Alexander Ypsilantis, a Phanariote who held the position of aide-de-camp to Alexander but who lacked the political experience of Kapodístrias, that the leadership was offered.

Like Rigas Velestinlis, the conspirators were hoping for the support of the Romanians and Bulgarians, but there was little enthusiasm for the project on the part of the other Balkan peoples, who were inclined to view the Greeks, with their privileged position in the Ottoman Empire and their enthusiasm for the ecclesiastical and cultural Hellenization of the other Balkan Christians, as scarcely less oppressive than the Turks.

Although unable to rely on the other Balkan peoples, the leadership of the conspiracy succeeded in exploiting the internal problems of the Ottoman Empire to its advantage. Sultan Mahmud II, who had ascended the throne in 1808, was bent on restoring the authority of the central government. To this end, in the winter of 1820, he launched an attack against Ali Paşa Tepelenë, a provincial warlord who, from his capital in Jannina, exercised control over large areas of mainland Greece. Although he nominally paid allegiance to the sultan, his virtual independence had for many years made him a thorn in the flesh of the Ottoman authorities. Taking advantage of the fact that large numbers of Ottoman troops were tied up in the campaign against Ali Paşa, Alexander Ypsilantis launched an attack from Russian territory across the Pruth River in March 1821, invoking the glories of ancient Greece in his call to arms from the Moldavian capital, Jassy. However, his campaign met with little success. He encountered no enthusiasm on the part of the supporters of Tudor Vladimirescu, who had risen against the oppression of the local Romanian boyars, or notables. Memories of Phanariote Greek oppression were altogether too vivid and recent. In June of 1821 Ypsilantis and his motley army were defeated at the battle of Dragatsani, and Ypsilantis was forced ignominiously to flee into Habsburg territory, where he died in captivity in 1828.

Revolt in the Peloponnese. Shortly after Ypsilantis' incursion into Moldavia, scattered violent incidents coalesced into a major revolt in the Peloponnese. With atrocities being committed by both sides, the Turks, very much in a minority, were forced to retreat to their coastal fortresses. The diversion of Ottoman forces for the attack on Ali Paşa, the element of surprise, and the military and naval skills on which the Greeks could draw, gave the Greeks an advantage in the early years of what proved a lengthy struggle.

The revolt caught the public imagination in western Europe, even if in the early years the reactionary governments of post-Napoleonic Europe were not prepared to countenance any disturbance of the existing order. Public sympathy in western Europe, however, was translated into more concrete expressions of support with the arrival in the Peloponnese of philhellene volunteers, the best-known of whom was the poet Lord Byron, who had traveled

Western sympathy

extensively in the Greek lands before 1821. The military contribution of the philhellenes was limited, and some became disillusioned when they discovered that Greek reality differed from the idealized vision of Periclean Athens in which they had been nurtured in their home countries. But, the philhellenic committees that sprang up in Europe and the United States raised money for the prosecution of the war and the succour of its victims, such as the survivors of the great massacre on Chios in 1822, immortalized by the French painter Eugène Delacroix.

Factionalism in the emerging state. At a very early stage in the fighting the question of the governance of the liberated territories arose. Initially no fewer than three provisional governments concurrently came into existence, while in 1822 a constitution, which by the standards of the day was highly democratic, was adopted with more than half an eye to securing the support of enlightened public opinion in Europe. A revised constitution was adopted in 1823, at which time the three local governments were unified in a central authority. However, unification did not bring unity. Feuding between rival groups culminated in outright civil war in 1824, prompting one chieftain, Makriyannis, to protest that he had not taken up arms against the Turks in order to end up fighting Greeks.

Civil war

Such factionalism derived from a number of causes. There was a basic tension between the *kodjabashis*, or notables, of the Peloponnese, who were anxious to ensure that they retained the privileged status they had held under the Ottomans, and the military element, associated with such klephtic leaders as Theodoros Kolokotronis, who sought recognition in terms of political power for their contribution to the war effort. The island shipowners, whose contribution to the prosecution of the war at sea was vital, likewise laid claim to a share of power, while the small intelligentsia argued for the adoption of liberal parliamentary institutions. To some degree the clash can be seen as a confrontation between Westernizers and traditional elites, to some degree as a clash between the military and civilian parties. The Westernizers, who were consciously nationalist and whose attitudes were expressed by their adoption of a Western lifestyle and Western clothing, wanted independent Greece to develop along the lines of a European state, with a regular army and with a curb on the traditional powers of the church. The traditional elites, on the other hand, tended to see the struggle in terms of a religious crusade against the Muslims, and their national consciousness was less fully articulated. Anxious to maintain the power and privileges they had enjoyed before the struggle began, they were chiefly concerned with substituting the oligarchy of the Turks with their own.

The insurgents could ill afford internecine fighting. Mahmud II had by this time forged an alliance with his nominal subject, Muḥammad 'Ali, the ruler of Egypt, and his son Ibrahim Pasha, who were promised lavish territorial rewards in return for their assistance in suppressing the revolt. From early 1825 Ibrahim Pasha engaged in a bitter war with the insurgents. As their initially favourable military position deteriorated, the insurgents looked increasingly for salvation to the Great Powers, which, from a combination of mutual suspicion as to each other's objectives and concern at the damage being done to their commercial interests, gradually moved toward a more interventionist position.

In 1826, by the Protocol of St. Petersburg, Britain and Russia committed themselves to a policy of mediation, to which France became a party through the Treaty of London of 1827. A policy of "peaceful interference," as the British prime minister Lord Canning described it, culminated in the not wholly planned destruction of the Turco-Egyptian fleet by a combined British, French, and Russian fleet at the Battle of Navarino in October 1827, the last great naval battle of the age of sail. This intervention by the Great Powers was instrumental in ensuring that some form of independent Greece came into existence, although its precise borders, which ran from Arta in the west to Volos in the east, took some years to negotiate. This process was overseen by Count Ioánnis Kapodistrias, who was elected the first president of Greece by the Assembly of Troezene, which in 1827 enacted the third constitution of the independence period. Besides overseeing the negotiation of the boundaries of the new state, in which his extensive diplomatic experience in the Russian imperial service was employed to the full, Kapodistrias was also fully engaged in trying to create the infrastructure of a state in a country that had been ravaged by a vicious and destructive war. Schooled as he was in the traditions of Russian autocracy, Kapodistrias chafed under the provisions of the 1827 constitution, which, like its predecessors, was a remarkably liberal document, and he abrogated it. His paternalist and authoritarian style of government offended a number of key elements in the power structure of the embryonic Greek state. Growing unrest culminated in his assassination in Nauplia, the provisional capital, in October 1831.

The Battle of Navarino

BUILDING THE NATION, 1832–1913

Greece's existence as an independent state gained formal recognition in the treaty of 1832 between Bavaria and the Great, or "Protecting," Powers. Significantly, the Greeks themselves were not party to the treaty. Greece now, formally at least, became a sovereign state, and the Greeks were thus the first of the subject peoples of the Ottoman Empire to gain full independence. But the state contained within its borders scarcely one-third of the Greek populations of the Middle East, and the struggle to expand the nation's borders was to dominate the first century of independent statehood. Only in 1947, with the incorporation of the Dodecanese Islands (a group of islands off the southwestern coast of Turkey hitherto under Italian rule), were Greece's present borders established.

Moreover, the sovereignty of the small Greek state was not absolute. The Protecting Powers had determined that Greece should be a monarchy, and they retained certain ill-defined rights of intervention.

Greece under Otto of Wittelsbach. The Great Powers had chosen Otto of Wittelsbach, the 17-year-old son of King Louis I (Ludwig) of Bavaria, as king of Greece. Because Otto was still a minor, the Great Powers determined that until he came of age the country was to be ruled by three Bavarian regents, while the army was to be composed of Bavarians. The period of the "Bavarokratia," as the regency was termed, was not a happy one, for the regents showed little sensitivity to the mores of Otto's adopted countrymen and imported European models wholesale without regard to local conditions. Thus the legal and educational systems were heavily influenced by German and French models, as was the church settlement of 1833, which ended the traditional authority of the ecumenical patriarch and subjected ecclesiastical affairs to civil control.

The "Bavarokratia"

Even after the formal ending of the regency in 1835, the Bavarian presence remained strong and was increasingly resented by those who had fought for independence and who now felt cheated of the fruits of victory. A further source of frustration for some was Otto's failure to grant a constitution, as had been provided for in the negotiations preceding independence. Despite the absence of a constitution, however, political parties of a kind came into existence; the "British," "Russian," and "French" parties were associated, as their names suggest, with the diplomatic representatives of the Great Powers, and their main appeal was strong personalities rather than well defined ideologies.

Toward the end of the decade of the 1830s there was growing discontent with Otto's rule. There was no sign of his conceding a constitution; Bavarians were still influential; his marriage to Queen Amalia had not produced an heir; the king remained a Roman Catholic in an Orthodox country with a strong anti-Catholic tradition; and much of the country's revenues were being expended in servicing the loan granted on independence by the Protecting Powers.

These various strands of discontent coalesced in the military coup of September 1843. Virtually bloodless, but on this occasion manifestly reflecting the popular will, the coup was the first of many military interventions in the political process. Otto was forced to grant a constitution (promulgated in 1844), which was a liberal document by

The constitution of 1844

the standards of the day, providing for virtually universal manhood suffrage (although women were barred from voting until as late as 1952). However, Otto, in concert with his wily prime minister, Ioannis Kolettis, was able to subvert the spirit of the new constitution by establishing a kind of parliamentary dictatorship. Moreover, the attempt to graft a liberal constitutional democracy onto an essentially premodern, traditional society that had evolved in quite a different fashion from those of western Europe gave rise to tensions both within the political system and in the relations between state and society which have continued until modern times. *Rouspheti* (the reciprocal dispensation of favours), patronage, manipulation, and, at times, outright force continued to characterize the politics of the postconstitutional period.

The Great Idea. It was during the debates that preceded the promulgation of the 1844 constitution that Kolettis first coined the expression the "Great Idea" (Greek: Megali Idea). This was a visionary nationalist aspiration that was to dominate foreign relations and, indeed, to a significant extent, to determine the domestic politics of the Greek state for much of the first century of its independent existence.

If the expression was new in 1844, the concept was deeply rooted in the Greek popular psyche, nurtured as it was by the prophecies and oracles that had kept alive hopes of eventual emancipation from the Turkish yoke during the dark centuries of the Tourkokratia. In essence, the Great Idea envisaged the restoration of the Christian Orthodox Byzantine Empire, with its capital once again established in Constantinople, which would be achieved by incorporating within the bounds of a single state all the areas of Greek settlement in the Middle East. Besides the Greek populations settled over a wide area of the southern Balkan Peninsula, there were extensive Greek populations in the Ottoman capital, Constantinople (Istanbul) itself; around the shores of the Sea of Marmara; along the western littoral of Asia Minor, particularly in the region of Smyrna (İzmir); in central Anatolia (Cappadocia), where much of the Greek populace was Turkish-speaking but employed the Greek alphabet to write Turkish; and in the Pontus region of northeastern Asia Minor, whose geographic isolation had given rise to a form of Greek barely intelligible elsewhere in the Greek world.

The Great Idea, the liberation by the Greek state of the "unredeemed" Greeks of the Ottoman Empire, was to be achieved through a combination of military means—an ambitious objective indeed for a state with such limited resources—and a far-reaching program of educational and cultural propaganda aimed at instilling a sense of Hellenic identity in the very large Greek populations that remained under Ottoman rule. The University of Athens (founded 1837) attracted people from all parts of the Greek world to be trained as students and apostles of Hellenism.

Greece hoped to profit from the Crimean War (1854–56) fought between Russia (as stated above, the only sovereign Orthodox power and looked upon for so long as the most likely liberator of the Greeks) and the Ottoman Empire and its British and French allies. However, Greek neutrality in the conflict was enforced by a British and French occupation of Piraeus, the port of Athens; this was one of several interventions in Greece's internal affairs by the Protecting Powers that made light of Greece's sovereign status.

King Otto's enthusiasm for the Great Idea at the time of the Crimean War was popular with his subjects, but during the 1850s there was renewed discontent. The manipulation of the 1844 constitution had alienated a younger generation of politicians who had had no direct experience of the war of independence. Moreover, Otto had still not converted to the Orthodox church and still had no heir. The king was driven into exile following a coup in 1862; throughout his remaining days in exile in Bavaria he demonstrated an affection for his subjects that was largely unreciprocated.

Reform, expansion, and defeat. The downfall of King Otto obliged the Great Powers to search for a new sovereign who could not be drawn from their own dynasties. Their choice was a prince of the Danish Glücksburg family, who reigned as King George I of the Hellenes from 1863 to 1913; thereafter the Glücksburg dynasty reigned intermittently until the 1974 referendum rejected the institution of monarchy. To mark the beginning of the new reign, Britain ceded to Greece the Ionian Islands, over which it had exercised a protectorate since 1815. This represented the first accession of territory to the Greek state since independence.

Political modernization. A new constitution in 1864 amplified the democratic freedoms of the 1844 constitution, although the sovereign retained substantial, if vaguely defined, powers in foreign policy. However, the realities of politics remained much as before, with numerous elections and even more frequent changes of administration, as politicians formed impermanent coalitions, jockeying for power in the disproportionately large parliament. It was not until 1875 that a decisive step was taken toward political modernization. In that year King George conceded that he would thenceforth entrust the government to the political leader enjoying the confidence of a majority of the deputies in parliament. During the last quarter of the 19th century the kaleidoscopic coalitions of earlier years gave way to a two-party system, in which power alternated between two men, Kharilaos Trikoúpis and Theodoros Deliyannis. Trikoúpis represented the modernizing, Westernizing trend in politics, while Deliyannis was a political boss in the traditional mold, with no real program other than overturning the reforms of his archrival Trikoúpis. Believing the modernization of the political system and economic development to be the essential preconditions of territorial expansion, Trikoúpis struggled to establish Greece's credit-worthiness in international markets and encouraged the country's hesitant steps in the direction of industrialization. He also promoted infrastructural projects such as road building, railway construction, the building of the Corinth Canal, and the draining of Lake Kopaïs in Thessaly. Such measures, however, and also Trikoúpis' concurrent efforts to modernize the country's armed forces, had to be paid for, and the increased taxation they entailed proved an easy target for a populist demagogue such as Deliyannis. Deliyannis was able to court further popularity by advocating an aggressive policy toward the Ottoman Empire, but his belligerence was to have disastrous economic consequences.

Trikoúpis and Deliyannis

Extension of Greek borders. If Britain had hoped to dampen irredentist enthusiasm by ceding the Ionian Islands, it was sorely mistaken. The continuing agitation in the "Great Island" of Crete for union with the Greek kingdom, which erupted in periodic uprisings, caused inevitable friction in relations with the Ottoman Empire, as did Greece's rather maladroit effort to exploit the latter's discomfiture in the great Middle Eastern crisis of 1875–78, which gave rise to a war between Russia and the Ottoman Empire. The Great Powers, meeting in Berlin in 1878, besides cutting down to size the "Big Bulgaria" that had arisen from the conflict, pressed the Ottoman government to cede the rich agricultural province of Thessaly and a part of Epirus to Greece. Thus, in 1881 the second extension of the territory of the independent state came about, like the first—the cession of the Ionian Islands—as a result of mediation by the Great Powers rather than of armed conflict. In 1878, again as part of the Berlin settlement, the island of Cyprus, with its largely Greek population, came under British administration, while formally remaining under Ottoman sovereignty. The island was annexed by Britain in 1914 after the Ottoman Empire had entered the war on the side of the Central Powers, becoming a crown colony in 1925.

Rectification of frontiers. The incorporation of Thessaly brought the northern frontier of Greece to the borders of Macedonia, which, with its mixed population of Greeks, Bulgarians, Serbs, Albanians, Turks, Vlachs, and Gypsies, was a byword for ethnic complexity. It also brought Greece into contention with Serbia and Bulgaria, all of which cast covetous eyes over Macedonia, which remained under Ottoman rule. Initially, the contest was conducted by means of ecclesiastical, educational, and cultural propaganda, but at the turn of the century rival guerrilla bands, financed by their respective governments (and supported by public

opinion), sought to achieve by terror what they could not achieve by more peaceful means.

While Trikoúpis argued for the strengthening of the state as the essential precondition of territorial expansion, Deliyannis showed no such caution. But his mobilization in 1885 in an attempt to exploit a crisis over Bulgaria resulted in the imposition of a naval blockade by the Great Powers, while his support for the insurgents in Crete in 1897 led to humiliating defeat in the Thirty Days' War with Turkey. Greece was forced to pay compensation and to accept rectifications of its frontier. Moreover, the repayment of its substantial external debts was to be overseen by an international financial commission, another humiliating erosion of its sovereignty.

Emigration. Military endeavours compounded serious economic problems, which had culminated in national bankruptcy in 1893. Economic difficulties were primarily responsible for the great wave of emigration, principally from the Peloponnese to the United States, that characterized the last decade of the 19th and the first decade of the 20th century. About one-sixth of the entire population participated in this great exodus. Very largely male, the early emigrants had little intention of settling permanently overseas; few, however, returned to their homeland, although most retained strong nostalgic ties to their birthplace. Migrant remittances to relatives in the old country thenceforward made a significant contribution to the country's balance of payments.

The early Venizélos years. The clear lesson of the 1897 war was that, however weakened the Ottoman state might be, Greece, notwithstanding the impassioned rhetoric of nationalist politicians, was in no position to engage in single-handed military confrontation. Allies and the reinvigoration of the ramshackle state and economy were the necessary preconditions of a successful military challenge. The latter came about under the inspired leadership of Eleuthérios Venizélos, who had made his mark in the politics of his native Crete where an autonomous regime had been established in the aftermath of the 1897 war. A charismatic figure who was adored and execrated in equal measure, Venizélos dominated Greek politics during the first third of the 20th century.

The Goudí coup. Venizélos was projected from the provincial to the national stage as a consequence of a coup staged by the Military League, formed by disaffected army officers, from Goudí (at the outskirts of Athens) in 1909; this coup ushered in a persistent pattern of military involvement in politics during the 20th century. The conspirators demanded thoroughgoing reforms of both a nonmilitary and a military nature, the latter including the removal of the royal princes, who were held to favour the promotion of their own protégés, from the armed forces.

Venizélos' reformist program. The short-lived but forceful intervention of the military compelled the discredited political establishment to make way for Venizélos, who had not been compromised by involvement in the petty politics of the kingdom. In elections held in December 1910 Venizélos and his newly founded Liberal Party won more than four-fifths of the seats in parliament. His power legitimized through elections, Venizélos plunged into a wide-ranging program of constitutional reform, political modernization, and economic development, which he combined with an energetic espousal of the Great Idea. Some 50 amendments to the 1864 constitution were enacted; provision was made for land reform; innovations were made in the educational system; and legislation benefiting the working population was introduced. These moderately reformist policies inhibited the development of the powerful agrarian and socialist movements that developed elsewhere in the Balkans. British naval and French military missions were brought in to overhaul the armed forces. Venizélos' continuing political ascendancy was confirmed with a sweeping victory in elections held in 1912.

The Balkan Wars. The pessimism induced by the defeat of 1897 gave way to a period of optimism, in which Greece grandiosely saw itself as a power in the ascendant, poised to displace a declining Ottoman Empire as the leading power in the Middle East. When, in 1911, Italy attacked the Ottoman Empire (in the process occupying the largely Greek-populated Dodecanese islands), Greece, no less than the other Balkan states, wanted its share of the spoils from the ever more likely collapse of Ottoman rule in the Balkans. However, Greece's situation differed from that of its Balkan neighbours, whose populations were relatively compactly settled within the Balkan peninsula. The Greeks alone were widely dispersed throughout the Middle East and hence were vulnerable to Turkish reprisals in the event of a war. Nonetheless, Greece could scarcely stand aside from the complex of alliances being formed among the Balkan states. These culminated in October 1912 in the First Balkan War, with Greece, Serbia, Bulgaria, and Montenegro declaring war on the Ottoman Empire. In contrast to earlier Balkan crises, the Great Powers did not intervene, and the heavily outnumbered Ottoman forces were forced into rapid retreat. Within less than a month, Thessaloníki (Salonica), the most important port in the northern Aegean, coveted by Bulgaria as well as by Greece, was captured by Greek forces. In February 1913 Greek forces took Ioánnina, the capital of Epirus. Meanwhile the Greek navy rapidly occupied the Aegean islands still under Ottoman rule.

The First Balkan War

The Balkan alliance was always a somewhat fragile affair in view of rivalries over Macedonia. Bulgaria, in particular, felt that its sacrifices had been in vain and turned against its erstwhile allies Greece and Serbia. This brief Second Balkan War (June to July 1913) led to the Treaty of Bucharest (August 1913), in which Bulgaria was forced to acknowledge the acquisition by Greece and Serbia of the lion's share of Macedonia. At the same time the formal union of Crete with the kingdom was recognized, although Greek hopes for the annexation of northern Epirus, with its large Greek population, were thwarted when the region was incorporated into the newly independent Albania.

The expansion of Greece's territories in the First and Second Balkan Wars had been truly breathtaking. Its land area had increased by some 70 percent, and so had its population (from 2.8 million to 4.8 million), but by no means were all of its newly acquired citizens ethnic Greeks. Indeed, in the city of Thessaloníki the largest single element in the city's population comprised Sephardic Jews, the descendants of the Jews expelled from Spain in 1492, who continued to speak Spanish. Elsewhere in "New Greece," as the recently acquired territories came to be known, there were substantial Slavic, Muslim (mainly Turkish), Vlach, and Gypsy populations. Like the Jews, many of these populations did not look upon the Greeks as liberators. The integration of "New" with "Old" Greece, the conservative core of the original kingdom, was not to be an easy process, but the problems it created did not emerge until much later.

At the conclusion of hostilities, Greece was gripped by euphoria. Under the charismatic leadership of Venizélos, the irredentist aspirations enshrined in the Great Idea appeared to be within reach. When King George I died at the hands of a deranged assassin in March 1913, there were demands that his successor, Crown Prince Constantine, be crowned not Constantine I (as he was) but Constantine XII to symbolize continuity with Constantine XI Palaeologus, the last emperor of Byzantium.

GREEK HISTORY SINCE WORLD WAR I

From National Schism to dictatorship. The dynamism and sense of national unity that had characterized the early Venizélos years gave way to rancour and vindictiveness that was to poison the country's political life throughout World War I and the interwar period. Greece was riven by the "National Schism," a division of the country into irreconcilable camps supporting either King Constantine I or his prime minister, Venizélos. The immediate grounds for tension were differences between the king and the prime minister as to Greece's alignment during World War I, although there were deeper causes underlying the split. The king advocated neutrality, while Venizélos was an enthusiastic supporter of the Triple Entente—Britain, France, and Russia—which he regarded as the alliance most likely to favour the implementation of Greece's remaining irredentist ambitions. The entente had, indeed, in an effort to lure Greece into the war, held out the attractive

The question of alignment

prospect of territorial gains for Greece at the expense of Turkey, which had aligned itself with the Central Powers. Increasingly bitter disagreements between king and prime minister resulted in the latter twice resigning in 1915, despite a convincing electoral victory.

The breach between the two became irrevocable when Venizélos in October of 1916 established a rival government in Thessaloniki, which, like most of "New Greece," was passionately loyal to Venizélos. In June 1917 the entente allies ousted King Constantine and installed Venizélos as prime minister of a formally united but bitterly divided Greece. Venizélos duly brought Greece, hitherto neutral, into the war on the side of the entente. Naturally, he expected to reap the reward for his loyalty at the Paris Peace Conference. In May of 1919 Greece was permitted to land troops in Smyrna (İzmir), the major port city in Asia Minor, with its large Greek population, and Greece was a major beneficiary of the Treaty of Sèvres of August 1920, the peace treaty with the defeated Ottoman Empire. However, for the Turkish nationalists, galvanized by the leadership of Mustafa Kemal (Atatürk), the treaty was from the outset a dead letter and the Greek landings a challenge they were prepared to meet.

In November 1920 Venizélos, to universal surprise, was defeated in elections, and the exiled King Constantine I was restored to his throne after a bogus plebiscite to the manifest displeasure of Britain and France. Meanwhile, the military situation in Asia Minor steadily deteriorated; a Turkish nationalist offensive in August/September 1922 resulted in a dramatic rout of the Greek armies in Asia Minor. Much of Smyrna was burned, and many Greeks and Armenians were killed. Tens of thousands of destitute Greek refugees fled to the kingdom. Thus ended a 2,500-year Greek presence in Asia Minor and with it the elusive vision of the Great Idea.

A military junta seized power in 1922 as King Constantine abdicated. Five royalist politicians and the deranged commander of the Asia Minor forces were tried and executed on a charge of high treason, although there was no evidence of deliberate treachery. The "Trial of the Six" was to poison the climate of interwar politics, exacerbating the already bitter feud between the supporters of Venizélos and of the monarchy.

Exchange of populations

At a peace conference in Lausanne an exchange of populations between Greece and the newly established Turkish Republic was agreed upon. The criterion employed was religion, one consequence of which was the exchange of many tens of thousands of Turkish-speaking Orthodox Christians for Greek-speaking Muslims. The ecumenical patriarchate was allowed to remain in Constantinople, as were the Greek inhabitants of that city and of the two islands, Imbros (now Gökçe) and Tenedos (Turkish: Bozca), which straddled the entrance to the strategically sensitive Dardanelles. In return, the Muslims of Greek Thrace were allowed to remain in situ. An influx of some 1.3 million refugees (including significant numbers from Russia and Bulgaria) strongly tested the social fabric of a country prostrated by some 10 years of intermittent war. Nonetheless, leaving aside the prejudice they encountered on the part of the indigenous population, the process of their integration into Greek society was remarkably successful. The economy, benefiting from the entrepreneurial skills of the refugees, experienced a significant degree of industrialization during the interwar period. The remaining large estates were broken up to provide smallholdings for the newcomers, and rural Greece became a society of peasant smallholders, which made for social stability albeit not for economic efficiency. The majority of the refugees were settled in the territories of "New Greece," thereby consolidating the area's "Greekness." Although refugees were disproportionately represented in the leadership of the newly founded Greek Communist Party (KKE), they by and large remained intensely loyal to Venizélos. Their vote was clearly instrumental in the formal establishment of a republic in 1923, shortly after the departure of King George II, who had briefly succeeded to the throne following his father's abdication in 1922. Indeed, the refugees and the army acted as the arbiters of political life during the interwar period.

In 1928 Venizélos made a political comeback, two years after the downfall of the short-lived military dictatorship headed by General Theodoros Pangalos in 1925–26. Although Venizélos initiated a good-neighbour policy with Italy and Greece's Balkan neighbours and brought about a remarkable rapprochement with Turkey, his government was knocked off course by the repercussions of the Stock Market Crash on Wall Street in 1929. Because Greece was dependent on the export of agricultural products such as olive oil, tobacco, and currants and on migrant remittances, it was severely affected by the slump in world trade.

Moreover, after four years of relative stability, politics reverted to the confusion of the early 1920s. When the anti-Venizélists won the 1933 elections, Colonel Nikolaos Plastiras, a staunch supporter of Venizélos and the mastermind behind the 1922 coup, sought to restore Venizélos to power by force. His coup was unsuccessful and was shortly afterward followed by an attempt on Venizélos' life. The political arena was once again polarized between supporters of Venizélos and of the monarchy. Fear of a royalist restoration lay behind another attempted coup by Venizelist officers in March 1935. Because of his proven involvement on this occasion, Venizélos was forced into exile in France, where he died shortly afterward, but not before he had urged his supporters to effect a reconciliation with the king.

The royalists were the main beneficiaries of the abortive 1935 coup, in the aftermath of which King George II had been restored to his throne, following a distinctly dubious plebiscite. Like Venizélos in exile, the king on his return to Greece was in a conciliatory mood. However, elections held under a system of proportional representation in January 1936 produced a deadlock between the two main parliamentary blocs, the Venizelists and the royalists. Both blocs engaged in secret negotiations with the communists, hitherto an insignificant force, who now, with 15 seats in the 300-seat parliament, held the balance of power.

The Metaxas regime and World War II. Public disillusionment with the endless intrigues of the political world, which had been growing apace in the preceding years, was exacerbated when the news leaked that the main political blocs were secretly negotiating with the communists. When the nonpolitical figure appointed by the king to head a caretaker government charged with overseeing the elections died, he was replaced as prime minister by General Ioannis Metaxas, a marginal figure on the far right of the political spectrum. Metaxas exploited labour unrest and a threatened general strike to persuade the king on Aug. 4, 1936, to suspend key articles of the constitution. Although the suspension purported to be temporary, parliament did not reconvene until 10 years later.

The Metaxas dictatorship

Backed by the army and tolerated by the king, the Metaxas dictatorship lasted four and a half years. The dictator, whose paternalistic style was signaled by the adoption of titles such as "National Father," "First Peasant," and "First Worker," shared the loathing of parliamentary democracy, liberalism, and communism characteristic of German Nazism and Italian Fascism, but the "Regime of the Fourth of August 1936" altogether lacked their dynamism. The Metaxas regime was neither aggressive nor racist, nor did it seek alliances with the European dictatorships. On the contrary, with the support of the king, Metaxas strove to maintain the country's traditional alignment toward Britain. The dictator, however, endeavoured to recast the Greek character in a more disciplined mode, invoking the values of ancient Greece and, in particular, of the Spartans. He furthermore sought to fuse them with the values of the medieval Christian Empire of Byzantium, thus fashioning what he pompously described as the "Third Hellenic Civilization."

At the outbreak of World War II Metaxas tried to maintain neutrality, but Greece was increasingly subject to pressure from Italy, whose dictator Benito Mussolini sought an easy military triumph to match those of his ally Adolf Hitler. A series of provocations culminated in the delivery of a humiliating ultimatum on Oct. 28, 1940. Metaxas, reflecting the mood of the entire nation, rejected this without discussion. The Italians immediately invaded Greece from Albania, which they had occupied 18 months

previously. But any hopes of a lightning military triumph were rapidly dashed. Within weeks not only had the Italians been driven from Greek territory, but Greek forces had pushed on to occupy much of what the Greeks term "Northern Epirus," the area of southern Albania with a substantial Greek minority.

While accepting token British military aid, Metaxas, until his death in January 1941, was anxious to avoid provoking German intervention in the conflict. However, his successor agreed to accept a British expeditionary force as it became apparent that Hitler's aggressive designs extended to the Balkans. The combined Greek and British forces, however, were able to offer only limited resistance when the German juggernaut rolled across the borders on April 6, 1941. By the beginning of June the country was overrun and subject to a harsh tripartite German, Italian, and Bulgarian occupation. King George II and his government-in-exile fled to the Middle East. The requisitioning of food stocks resulted in a terrible famine during the winter of 1941–42, in which as many as 100,000 people died. In 1943 virtually the entire Jewish population was deported to death camps in Germany. A devastatingly high rate of inflation added to the miseries and humiliations of everyday life.

Almost from the outset of the occupation, acts of resistance were recorded. These took a more systematic form after the Communist Party in September 1941 founded the National Liberation Front (EAM), whose military arm was known under the initials ELAS. Although the communists had been a marginal force during the interwar period, EAM/ELAS became the largest resistance organization. Other groups came into being, the most important of which, the National Republican Greek League (EDES), opposed, as did EAM/ELAS, the return of the king upon liberation. With the support of a British military mission, the guerrillas engaged in some spectacular acts of resistance, the most notable of which was the destruction in November 1942 of the Gorgopotamos viaduct, which carried the railway line from Thessaloníki to Athens.

The National Liberation Front

Just as during the War of Independence and World War I, so during this time of grave national crisis internecine strife divided the resistance organizations. Besides fighting the Axis occupation, they jockeyed for postwar power. During the winter of 1943–44 civil war broke out in the mountains of Greece between EAM/ELAS and the much smaller EDES, which, however, enjoyed the support of the British authorities, who had become increasingly alarmed at the prospect of a postliberation seizure of power by the communists.

Not for the first time in Greek history, the country's fate was to be determined by the Great Powers. The British prime minister, Winston Churchill, eager to see King George II restored to his throne, engaged in the summer and autumn of 1944 in some high-level negotiations with the Soviet leader, Joseph Stalin, trading Russian predominance in postwar Romania for British predominance in Greece. True to the spirit of this deal, it would seem, Stalin gave no encouragement to the Greek communists to make a bid for power in the autumn of 1944 as the Germans began their withdrawal, even though by this time they were by far the most powerful force in occupied Greece.

The confrontation was only postponed, however, for bloody fighting broke out in Athens in December 1944 between ELAS and the small British force that had accompanied the Greek government on its return from exile on October 18. It was a sign of Churchill's obsession with the crisis in Greece that he flew to Athens on Christmas Eve 1944 in an unsuccessful attempt to resolve it. The British prime minister, however, was able to persuade King George II not to return to Greece pending a plebiscite on the monarchy and to accept the regency of Archbishop Damaskinos of Athens. A temporary respite in the struggle between left and right was achieved at the Varkiza conference in February 1945, which aimed at a political settlement of the crisis.

Fighting in Athens

Civil war and its legacy. The first elections since the fateful ones of 1936 were held in March 1946. These, however, were flawed, and, with the far left abstaining, resulted in a sweeping victory for the royalist right. In September a plebiscite issued in a vote for the return of King George II; he died within six months and was succeeded by his brother Paul. Against this background the country slid toward civil war as the far left was undecided as to whether to work within the political system or to make an armed bid for power.

The turning point toward civil war came with the establishment in October 1946 of a communist-controlled Democratic Army. In December 1947 the communists established a Provisional Democratic Government. Although heavily outnumbered, the communists were able, with the logistical support from the newly established communist regimes to the north, coupled with skillful use of guerrilla tactics, to control a wide area of northern Greece for a substantial period of time. Following the declaration of the Truman Doctrine in March 1947, which pledged support for "free peoples" in their fight against internal subversion, the tide gradually began to turn, as the United States, assuming Britain's former mantle as Greece's chief external patron, provided military equipment and advice. American intervention and the consequences of the break between Josip Broz Tito and Stalin, combined with factionalism and altered military tactics on the left, all contributed to the defeat of the communist guerrillas in the summer of 1949.

Greece emerged from the travails of the 1940s in a state of devastation. Nonetheless, if the post-civil-war political regime had a distinctly authoritarian hue, from the mid-1950s Greece underwent a rapid, if unevenly distributed, process of economic and social development, far surpassing its communist neighbours to the north in standard of living. The population of greater Athens more than doubled in size between 1951 and 1981, and by the early 1990s about one-third of the entire population was concentrated in the area of the capital. However, if urbanization progressed quickly and living standards rose rapidly, the country's political institutions failed to keep pace with rapid change. Following a brief centrist interlude, the right maintained a firm grip on power between 1952 and 1963 and was none too scrupulous about the means it employed to retain it.

By the early 1960s, however, the electorate (which now included women) had become increasingly disenchanted with the repressive legacy of the civil war and looked for change. This was offered by Georgios Papandreou, whose Centre Union Party secured a sweeping victory in 1964. Yet the promise of reform and modernization was thwarted as, against a background of renewed crisis in Cyprus, groups within the army conspired to subvert the country's democratic institutions. A guerrilla campaign in Cyprus, fought from the mid-1950s onward with tenacity and ruthlessness by the Greek-Cypriot general Georgios Grivas, had resulted in 1960 in the British conceding not the union with the Greek state sought by the overwhelming Greek-Cypriot majority on the island but independence. However, within three years the elaborate power-sharing arrangements between the Greek majority and the Turkish minority on the island had broken down.

During the civil war and after, Greece's armed forces had come to look upon themselves not only as the country's guardians against foreign aggression but also as its defenders against internal subversion, of which they were to be the final judge. They increasingly viewed Georgios Papandreou as a stalking horse for his much more radical American-educated son, Andreas Papandreou, who had returned to Greece and joined his father's government.

On April 21, 1967, middle-ranking officers, led by Colonel Georgios Papadopoulos, launched a coup designed to thwart an expected Centre Union victory in elections planned for May of that year. The conspirators took advantage of a prolonged political crisis, which had its origins in a dispute between the young King Constantine II, who had succeeded his father, King Paul, to the throne in 1964, and his septuagenarian prime minister, Georgios Papandreou. Alternating between policies that were heavy-handed and absurd, the "Colonels," as the military junta came to be known, misruled the country between 1967 and 1974. In December 1967, after a failed countercoup, King Constantine went into exile, with Papadopoulos as-

The Colonels

suming the role of regent. In 1973, following student protests, which were violently suppressed, Papadopoulos was toppled from within the junta, to be replaced by the even more repressive brigadier general Demetrios Ioannidis, the head of the much-feared military police.

In July 1974, in the wake of an increasingly bitter dispute between Greece and Turkey over oil rights in the Aegean Sea, Ioannidis, seeking a nationalist triumph, launched a coup to depose Makarios III, the archbishop and president of Cyprus since 1960. Makarios survived, but the coup triggered the invasion of the northern part of the island by Turkey, which, together with Britain and Greece, was a guarantor of the 1960 constitutional settlement. The Turkish army occupied almost 40 percent of the land area of the island, despite the fact that the Turkish population numbered less than 20 percent. Ioannidis' response to the Turkish invasion was to mobilize for war with Turkey. The mobilization proved chaotic, however, and the regime, bitterly unpopular domestically and totally isolated diplomatically, collapsed in complete disarray.

Restoration of democracy. Konstantinos Karamanlis, a conservative politician, was summoned back from self-imposed exile in France to restore democracy and rebuild a country ravaged by seven years of brutal and inefficient military rule. This he accomplished with signal success. He defused the threat of outright war with Turkey, ensured that the army returned to the barracks, and, acknowledging the way in which opposition to the junta had brought together politicians of all political backgrounds, legalized the Communist Party, which had been outlawed in 1947. He moved rapidly to legitimize his power through elections held in November 1974, in which he secured a sweeping victory. In December of the same year, a referendum on the future of the monarchy resulted in a 69 percent vote against the monarchy and the return of King Constantine.

Karamanlis' second premiership lasted from 1974 until 1980, when he was elected president. By this time he had achieved his main objective, early membership in the European Community, which Greece joined in January 1981. His failure to counter the populist appeal of Andreas Papandreou's Panhellenic Socialist Movement (PASOK) resulted in a stunning electoral victory for PASOK in 1981.

The smooth transfer of power from a right-wing government that had ruled for much of the postwar period to a radical (at the level of rhetoric at least) socialist one appeared to indicate that Greece's newly reestablished democratic institutions were firmly in place. During nearly a decade of socialist rule, Papandreou's promises of change and a dramatic reorientation in the country's domestic politics and external relations were not fulfilled. Ambitious plans to "socialize" key sections of industry failed to materialize, and the attempt to create a welfare state could be sustained only by enormous borrowing. Important reforms were, however, introduced in family law, and society was liberalized in other respects. It was testimony to Papandreou's ability to articulate both the aspirations and frustrations of much of the electorate that, despite a poor economic record and amid accusations of large-scale corruption in the higher reaches of his party, the conservative New Democracy Party only barely regained power in 1990. The new government, with Konstantinos Mitsotakis as prime minister, was committed to a policy of economic liberalism and the diminution of the powers of the state, but the problems that confronted it were formidable. The policies of rigid economics introduced by Mitsotakis and, in particular, proposals for the privatization of the large state sector were unpopular with much of the electorate.

In 1993 Papandreou's PASOK returned to power with a share of the vote only marginally smaller than it had received at the time of its electoral triumph in 1981. However, the underlying economic and infrastructural problems facing Greece remained, and Papandreou was unable to halt the rising deficits, inflation, interest rates, and unemployment. With Papandreou in his mid-seventies and in failing health, PASOK was divided over who should be his successor. By January 1996, Papandreou was so incapacitated that he resigned; he died later that year. Konstantinos (Kostas) Simitis was elected the new prime minister by the PASOK parliamentary deputies.

Simitis, more a pragmatic reformist than an ideological socialist, tried to control state spending and even to privatize some state industries. Hoping to capitalize on his apparent popularity, he called for elections in September, barely defeating the New Democracy Party. In the late 1990s, Simitis's efforts to introduce some fundamental restructuring of Greece's economy—taxes, labour laws, privatization, and social benefits—had some success but also provoked resistance from many Greeks. Much of the impetus for reforming Greece's economy came from its membership in the European Union (EU). Although Greece initially failed to meet the requirements for joining the European Economic and Monetary Union, it adopted the euro, the EU's single currency, in 2001.

In its foreign relations, Greece continued to be preoccupied with disputes related to its EU membership. After years of resisting the establishment of an independent nation of Macedonia, Greece reluctantly agreed in September 1995 to recognize its existence under the name of The Former Yugoslav Republic of Macedonia. Relations between Albanians (both those in Albania and in Greece) and Greeks continued to be strained. When NATO bombed Serbia in 1999 in defense of the Kosovar Albanians, Greece refused to participate in the air attacks. Greece's centuries-old "cold war" with Turkey was further exacerbated by quarrels over Cyprus, oil-drilling rights in the Aegean, and airspace violations. However, earthquakes that devastated Turkey and Greece in the summer of 1999 led each country to provide rescue teams, resulting in a sudden thaw in their relations.

For later developments in the history of Greece, see the BRITANNICA BOOK OF THE YEAR. (R.R.M.C./J.S.Bo./Ed.)

BIBLIOGRAPHY

General works. All aspects of the country are treated in GLENN E. CURTIS (ed.), *Greece: A Country Study*, 4th ed. (1995). JOHN CAMPBELL and PHILIP SHERRARD, *Modern Greece* (1968), contains, besides useful historical surveys, valuable chapters on the Orthodox church, literature, and the economy, while paying attention to the values underpinning society. YORGOS A. KOURVETARIS (GEORGE A. KOURVETARIS) and BETTY A. DOBRATZ, *A Profile of Modern Greece: In Search of Identity* (1987), contains material on many aspects on contemporary Greece. A good source for readings on Greece is MARY IO CLOGG and RICHARD CLOGG (compilers), *Greece* (1980), a bibliography with more than 800 entries on some 30 subjects, with the majority of cited sources in English.

Physical and human geography. One of the most extensive works on the geography of Greece is ALFRED PHILIPPSON, *Die griechischen Landschaften*, 4 vol. (1950–58). H.C. DARBY *et al.*, *Greece*, 3 vol. (1944–45), produced by the Naval Intelligence Division of Great Britain, contains much material of value on physical and economic geography. J.L. MYRES, *Dodecanese*, 2nd ed. (1943), also produced by the Naval Intelligence Division, is a survey of the Dodecanese islands under Italian rule between 1912 and 1947.

Greece's geology is treated in a regional context in CLIFFORD EMBLETON (ed.), *Geomorphology of Europe* (1984), chapter 16; and DEREK V. AGER, *The Geology of Europe* (1980), chapters 15–16. PIERRE BIROT and JEAN DRESCH, *La Méditerranée et le Moyen-Orient*, vol. 2, *La Méditerranée Orientale et le Moyen-Orient* (1955), offers details on physical structure and brief treatments of climate and vegetation. Individual aspects of the landscape are dealt with in E.G. MARIOLOPOULOS, *An Outline of the Climate of Greece* (1961; originally published in Greek, 1953). J.R. MCNEILL, *The Mountains of the Mediterranean World: An Environmental History* (1992), includes the Pindus Mountains as one of the case studies. (C.D.S.)

Classic studies of Greece's people and customs include ERNESTINE FRIEDL, *Vasilika: A Village in Modern Greece* (1962); and J.K. CAMPBELL, *Honour, Family, and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community* (1964, reissued 1974). MICHAEL KENNY and DAVID I. KERTZER (eds.), *Urban Life in Mediterranean Europe* (1983), includes several essays on Greece, including a study of rural-urban migration. TIMOTHY WARE (KALLISTOS WARE), *The Orthodox Church*, rev. 2nd ed. (1997), is a clear and concise account of the history and theology of the predominant religion in Greece.

The economy is covered by A.F. FRERIS, *The Greek Economy in the Twentieth Century* (1986); and PERSEFONI V. TSALIKI, *The Greek Economy: Sources of Growth in the Postwar Era* (1991). Politics is dealt with in KEITH R. LEGG, *Politics in Modern Greece* (1969); and RICHARD CLOGG, *Parties and Elections in Greece* (1987).

The remarkable continuities in the Greek language are dis-

cussed in ROBERT BROWNING, *Medieval and Modern Greek*, 2nd ed. (1983). A comprehensive survey, beginning with the emergence in the 11th century AD of literature in a recognizably modern form of the language, is LINOS POLITIS (LINOS POLITÉS), *A History of Modern Greek Literature* (1973). (R.R.M.C.)

History. *Greece during the Byzantine period (c. AD 300–c. 1453):* JOHANNES KODER and FRIEDRICH HILD, *Hellas und Thessalia* (1976), provides a detailed regional historical and geographic survey and includes an extensive bibliography as well as a discussion of the historical and political evolution of the region.

General surveys of the history of the Byzantine world all include information dealing with Greece at the appropriate junctures. The most useful are GEORGE OSTROGORSKY (GEORGIJE OSTROGORSKI), *History of the Byzantine State*, 2nd ed. (1968, reissued 1980; originally published in German, 1940); and D.M. HUSSEY, D.M. NICOL, and G. COWAN (eds.), *The Byzantine Empire*, 2nd ed., 2 vol. (1966–67), vol. 4 of *The Cambridge Medieval History*. A wealth of detail on the society and economy of the late Roman world, as well as on the provincial administration of the Greek regions, is provided by A.H.M. JONES, *The Later Roman Empire, 284–602: A Social Economic and Administrative Survey*, 2 vol. (1964, reprinted 1986). The transition from late Roman to early Byzantine structures, the fate of urban society, and the effects of the disruptions of the 7th century are surveyed in J.F. HALDON, *Byzantium in the Seventh Century: The Transformation of a Culture*, rev. ed. (1997), with detailed discussion of a number of fundamental developments. Society and economy in the later period are treated in ALAN HARVEY, *Economic Expansion in the Byzantine Empire, 900–1200* (1989), for the period to the Fourth Crusade; and ANGELIKI E. LAIOU-THOMADAKIS, *Peasant Society in the Late Byzantine Empire* (1977), for the period from about 1204 until the end of the empire. MICHAEL F. HENDY, *Studies in the Byzantine Monetary Economy, c. 300–1450* (1985), presents a detailed collection of surveys of the physical geography, land use, and settlement patterns of the Balkans (as well as other regions of the empire), together with a discussion of the nature of the Byzantine economy, the fiscal administration, and related topics. NICOLAS OIKONOMIDÈS, *Les Listes de préséance byzantines des IXe et Xe siècles* (1972), presents the evidence for the development of the middle Byzantine provincial, fiscal, and administrative structures that evolved in Greece during this period. The most accessible materials for demography and population are PETER CHARANIS, "On the Demography of Medieval Greece: A Problem Solved," *Balkan Studies*, 20 (2):193–218 (1979); and PETER CHARANIS (compiler), *Studies on the Demography of the Byzantine Empire* (1972), a collection of articles.

Works dealing specifically with Greece include APOSTOLOS E. VACALOPOULOS, *Origins of the Greek Nation*, trans. from Greek (1970); and NICOLAS CHEETHAM, *Mediaeval Greece* (1981), both of which provide excellent general accounts, the former in particular presenting both the political, socioeconomic, and ethnic-linguistic issues. English-language surveys of different regions are DONALD M. NICOL, *The Despotate of Epiros, 1267–1479* (1984); DENIS A. ZAKYTHINOS (DIONYSIOS A. ZAKYTHÈNOS), *Le Despotat grec de Morée*, 2 vol., rev. and augmented by CHRYSSA MALTÉZOU (1975); and MICHAEL ANGOLD, *A Byzantine Government in Exile: Government and Society Under the Lascarids of Nicaea, 1204–1261* (1974). Particular aspects of regional history are discussed in DAVID JACOBY, *Recherches sur la Méditerranée orientale du XIIe au XVe siècle: peuples, sociétés, économies* (1979); and PETER TOPPING, "The Morea, 1311–1364," and "The Morea, 1364–1460," in KENNETH M. SETTON (ed.), *A History of the Crusades*, vol. 3 (1975), pp. 104–166, all of which deal with social and economic as well as political and historical problems connected with the Latin/Frankish presence in Greece. The roles of the Vlachs and Albanians are examined by T.J. WINNIFRITH, *The Vlachs* (1987); and ALAIN DUCELLIER, *L'Albanie entre Byzance et Venise: Xe–XVe siècles* (1987). A useful and important survey of Byzantium and the Slavs, as well as the Vlachs and Albanians, is DIMITRI OBOLENSKI, *The Byzantine Commonwealth: Eastern Europe, 500–1453* (1971, reissued 1982). A basic reference work that deals with all the topics referred to, sometimes in detail, and that also includes further references is ALEXANDER P. KAZHDAN (ed.), *The Oxford Dictionary of Byzantium*, 3 vol. (1991). (J.F.H.)

Greece under Ottoman rule, 1453–1831: ARNOLD TOYNBEE, *The Greeks and Their Heritages* (1981), is a stimulating survey of the whole range of Greek history from prehistoric times to the present day. One of the few scholarly studies in English of the dark age of Greek history, between the fall of Constantinople and the capture of Crete, is APOSTOLOS E. VACALOPOULOS (APOSTOLOS E. VAKALOPOULOS), *The Greek Nation, 1453–1669: The Cultural and Economic Background of Modern Greek Society*, trans. from Greek (1976). An overview of the critical four centuries of Ottoman rule is contained in D.A. ZAKYTHINOS (DIONYSIOS A. ZAKYTHÈNOS), *The Making of Modern Greece: From Byzantium to Independence*, trans. from Greek (1976). The crucial role of the church during the period is discussed in STEVEN RUNCIMAN, *The Great Church in Captivity: A Study of the Patriarchate of Constantinople from the Eve of the Turkish Conquest to the Greek War of Independence* (1968, reissued 1986). RICHARD CLOGG (ed. and trans.), *The Movement for Greek Independence, 1770–1821* (1976), illustrates the emergence of the Greek national movement through contemporary documents; while G.P. HENDERSON, *The Revival of Greek Thought, 1620–1830* (1970), focuses on the intellectual revival that preceded the outbreak of the war of independence in 1821. The war itself is covered in DOUGLAS DAKIN, *The Greek Struggle for Independence, 1821–1833* (1973); and the diplomacy of the period is analyzed in C.W. CRAWLEY, *The Question of Greek Independence: A Study of British Policy in the Near East, 1821–1833* (1930, reprinted 1973). The colourful story of the philhellene volunteers who fought alongside the insurgent Greeks is told by WILLIAM ST. CLAIR, *That Greece Might Still Be Free: The Philhellenes in the War of Independence* (1972). The independence movement is also traced in C.M. WOODHOUSE, *Capodistria: The Founder of Greek Independence* (1973), a study of the first president of Greece.

Greece since 1831: RICHARD CLOGG, *A Concise History of Greece* (1992), an illustrated survey, focuses mainly on the 19th and 20th centuries. DOUGLAS DAKIN, *The Unification of Greece, 1770–1923* (1972), is a detailed study of the gradual expansion of the Greek state. E.S. FORSTER, *A Short History of Modern Greece, 1821–1956*, 3rd ed. rev. and enlarged (1958, reprinted 1977), is a basic survey. *Journal of Modern Greek Studies* (semiannual), published by the Modern Greek Studies Association, is (since 1983) the premier publication for scholarship on modern Greece. The early years of King Otto's reign are studied in considerable detail in JOHN A. PETROPOULOS, *Politics and Statecraft in the Kingdom of Greece, 1833–1843* (1968). CHARLES K. TUCKERMAN, *The Greeks of To-day*, 3rd ed. rev. and corrected (1886), is a perceptive account of mid-19th century Greece written by the first U.S. minister to Greece. The Goudi coup of 1901, the first of many military interventions in the political process in the 20th century, is the subject of S. VICTOR PAPACOSMA, *The Military in Greek Politics* (1977). The meddling of British and France in Greece's internal affairs during the First World War is treated by GEORGE B. LEON, *Greece and the Great Powers, 1914–1917* (1974); while MICHAEL LLEWELLYN SMITH, *Ionian Vision: Greece in Asia Minor, 1919–1922* (1973, reissued 1998), thoroughly treats the disastrous Anatolian entanglement. GEORGE TH. MAVROGORDATOS, *Stillborn Republic: Social Coalitions and Party Strategies in Greece, 1922–1936* (1983), provides an indispensable guide to the complex politics of the interwar period. The impact of the Axis occupation is investigated by MARK MAZOWER, *Inside Hitler's Greece: The Experience of Occupation, 1941–44* (1993, reissued 1995). A critical decade of foreign occupation, resistance, and civil war is the subject of C.M. WOODHOUSE, *The Struggle for Greece, 1941–1949* (1976, reissued 1979), while his *The Rise and Fall of the Greek Colonels* (1985) analyzes one of the consequences of the civil war, the military dictatorship of 1967–74. C.M. WOODHOUSE, *Karamanlis: The Restorer of Greek Democracy* (1982), traces the political career of the politician who oversaw the return to democracy. The whirlwind rise to power in 1981 of Andreas Papandreou and his PASOK party is the subject of MICHALIS SPOURDALAKIS, *The Rise of the Greek Socialist Party* (1988). THANOS VEREMIS (THANOS VEREMÈS), *The Military in Greek Politics: From Independence to Democracy* (1997), is a survey of the often strained relations between Greece's military and civilian powers. Books that focus on Greece since the 1960s include LORING M. DANFORTH, *The Macedonian Conflict: Nationalism in a Transnational World* (1995), an account of an issue that continues to vex Greeks; DAVID HOLDEN, *Greece Without Columns* (1972), which is blunt and opinionated but perceptive; THEODORE C. KARIOTIS (ed.), *The Greek Socialist Experiment: Papandreou's Greece, 1981–1989* (1992), a balanced collection of essays on this controversial period; DIMITRI CONSTAS and THEOFANIS G. STAVROU (eds.), *Greece Prepares for the Twenty-First Century* (1995); VAN COUFOUDAKIS, HARRY J. PSOMIADES, and ANDRE GEROLYMATOS (eds.), *Greece and the New Balkans: Challenges and Opportunities* (1999); and KOSTAS A. LAVDAS, *The Europeanization of Greece: Interest Politics and the Crises of Integration* (1997).

(R.R.M.C./J.S.Bo.)

Ancient Greek and Roman Civilizations

Ancient Greek and Roman civilizations is used here loosely, as an overall term referring to the Aegean and ancient Greek civilizations and the civilizations of the Etruscans and the ancient Italic peoples as well as the ancient Roman civilization. All of these flourished on the continents of Europe and Asia from the Neolithic Period (New Stone Age) to the decline of the Western Roman Empire in the 5th century AD and are treated here.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 912, and the *Index*.

The article is divided into the following sections:

Aegean civilizations 205
  Early Aegean civilizations 207
    Early cultures
    The Bronze Age
  The decline of the early Aegean civilizations 213
    The eruption of Thera (*c.* 1500) and the conquest of Crete (*c.* 1450)
    The end of the Bronze Age in the Aegean
    The people of the Aegean Bronze Age
Ancient Greek civilization 218
  The early Archaic period 218
    The post-Mycenaean period and Leftkandi
    Colonization and city-state formation
    Early Archaic Greek civilization
  The later Archaic periods 223
    The rise of the tyrants
    Sparta and Athens
    The world of the tyrants
  Classical Greek civilization 233
    The Persian Wars
    The Athenian empire
    The Peloponnesian War
    Greek civilization in the 5th century
  The 4th century 248
    To the King's Peace (386 BC)
    From 386 BC to Philip II of Macedon
    The rise of Macedon
    Alexander the Great
    Greek civilization in the 4th century
  Conclusion 263
Hellenism 264
  Political developments 264
    Alexander's successors
    The mid-3rd century
    The coming of Rome (225–133)
    The Greek world under the Roman Empire
  Hellenistic civilization 268
    Institutions and administrative developments
    Economic developments
    Cultural developments
    Science and medicine
    Philosophy
Ancient Italic peoples 272
  The Etruscans 272
    General considerations
    Language and writing
    Archaeological evidence
    Religion and mythology
    Origins
    Expansion and dominion
    Organization
    Crisis and decline
  Other Italic peoples 278
Ancient Rome 280
  Rome from its origins to 264 BC 281
    Early Rome to 509 BC
    Early centuries of the Roman Republic
    Roman expansion in Italy
  The middle republic (264–133 BC) 287
    The first two Punic Wars
    The establishment of Roman hegemony in the Mediterranean world
    Beginnings of provincial administration
  The transformation of Rome and Italy during the middle republic 293
    Citizenship and politics in the middle republic
    Culture and religion
    Economy and society
    Rome and Italy
  The late republic (133–31 BC) 299
    The aftermath of the victories
    The reform movement of the Gracchi (133–121 BC)
    The republic (*c.* 121–91 BC)
    Wars and dictatorship (*c.* 91–80 BC)
    The Roman state in the two decades after Sulla (79–60 BC)
    The final collapse of the Roman Republic (59–44 BC)
    The Triumvirate and Octavian's achievement of sole power
  Intellectual life of the late republic 307
  The early Roman Empire (31 BC–AD 193) 308
    The consolidation of the empire under the Julio-Claudians
    Growth of the empire under the Flavians and Antonines
    The empire in the 2nd century
  The later Roman Empire 323
    The dynasty of the Severi (AD 193–235)
    Religious and cultural life in the 3rd century
    Military anarchy and the disintegration of the empire (235–270)
    The recovery of the empire and the establishment of the dominate (270–337)
    The Roman Empire under the 4th-century successors of Constantine
    The eclipse of the Roman Empire in the West (*c.* 395–500) and the German migrations
Bibliography 337

Aegean civilizations

The adjective Aegean is traditionally used in reference to the Stone and Bronze Age civilizations that arose and flourished in the area of the Aegean Sea in the periods, respectively, about 7000–3000 BC and about 3000–1000 BC.

The area consists of Crete, the Cyclades and some other islands, and the Greek mainland, including the Peloponnese, central Greece, and Thessaly. The first high civilization on European soil, with stately palaces, fine craftsmanship, and writing, developed on the island of Crete. Later, the peoples of the mainland adapted the Cretan civilization to form their own, much as the Romans adapted the civilization of later Greece. The Bronze Age civilization of Crete has been called Minoan, after the legendary King Minos of Knossos, which was the chief city of the island throughout early times. The Bronze Age of the Cyclades is known as Cycladic, that of the mainland as Helladic, from Hellas, the Greek name for Greece. Early, middle, and late stages have been defined in each of these, with further subdivisions according to recognizable changes in the style of pottery and other products that are associated with each separate culture. The civilization that arose on the mainland under Cretan influence in the 16th century BC is called Mycenaean after Mycenae, which appears to have been one of its most important centres. The term Mycenaean is also sometimes used for the civilizations of the Aegean area as a whole from about 1400 BC onward.

Minoan, Cycladic, and Helladic civilizations

Dating of the Aegean Bronze Age. The dates that are suggested here are approximate and conventional. In a general way, they are based on correlations with Egypt, where, from the beginning of the Early Dynastic period (*c.* 2925 BC onward), a historical chronology can be established with a leeway of a few centuries and can be fixed within reasonably narrow limits after about 2000 BC. Bronze Age pottery from the Aegean has been found



Principal sites associated with Aegean civilizations.

in Egypt in contexts that are datable, and many Egyptian objects have been recovered on the island of Crete.

Two important landmarks are fragments of Cretan pottery from the town at Kahun in the Fayyūm, built for workers engaged in the construction of a pyramid for the pharaoh Sesostris II (ruled 1897–78), and a large quantity of Mycenaean pottery from the mainland found at Tell el-Amarna, site of Akhenaton's capital, and imported during his reign (*c.* 1350–34). Radiocarbon dates appear consistent with those based on correlations with Egypt. Objects found in 1982 in the Kaş-Ulu Burun shipwreck off the southern coast of Turkey, including the first known gold scarab of the Egyptian queen Nefertiti, reveal a tight web of interconnections in the later 14th century among Mycenaean Greece, Cyprus, Egypt, Palestine, Syria, and Africa.

History of exploration. The poems of Homer, which reflect an epic tradition that absorbed many changes occurring in warfare and society between the 15th and the 8th century BC, describe warriors employing bronze weapons and objects such as helmets plated with tusks of wild boar that went out of use before the end of the Aegean Bronze Age. Massive Bronze Age defense walls survived at Mycenae and elsewhere on the mainland; they were called Cyclopean because, according to Greek tradition, the Cyclopes had built them. Apart from these Cyclopean walls, virtually nothing was known about the Aegean Bronze Age before the middle of the 19th century, when in 1876 a German archaeologist, Heinrich Schliemann, discovered unplundered royal shaft graves at Mycenae. He thought that the men buried in them were the Greek heroes of Homer's siege of Troy. There are in fact many likenesses between Homer's descriptions and the armour, weapons, and war imagery found in these graves. The graves, spanning about 1600 to 1450 BC, contained princely gifts from an age when Greece, Crete, and Troy engaged in trade. Schliemann's discoveries led to intensive exploration of Bronze Age and earlier sites on the Greek mainland. On the island of Thera in 1866–67, before Schliemann, Ferdinand Fouqué, a French geologist, had already explored settlements of the Shaft Grave Period sealed in under a thick shroud of volcanic pumice and ash. He found houses, frescoes, pottery imported from as far as Cyprus, and well-preserved agricultural produce. Because Bronze Age Crete and Greece were not explored at the time, this important find lay fallow for a century.

Later in the 19th century, Christos Tsountas, a Greek archaeologist, dug cemeteries of earlier phases of the Bronze Age on other Cycladic islands and continued the work begun by Schliemann at Mycenae. At the end of the century, a British expedition excavated the important Bronze Age town of Phylakopi on Melos. When Crete eventually became independent of Turkish rule in 1898, attention was turned to Bronze Age sites there. In 1900 Arthur (later Sir Arthur) Evans, an English archaeologist, began to uncover the palace at Knossos, the largest Bronze Age centre of the island, discovering clay tablets with the first positive evidence for Bronze Age writing in the Aegean. Greek, American, French, and Italian excavators added further knowledge of the Cretan Bronze Age during the years that followed, and American and German expeditions opened

The work of Schliemann

Bronze Age writing

new sites on the mainland. Inscribed clay tablets in the script called Linear B, such as those found at Knossos in Crete at the turn of the century, were recovered in Messenia in 1939 by the American archaeologist Carl W. Blegen; others have since come to light at Mycenae and elsewhere on the mainland. The belief that the language of these tablets was a very archaic form of Greek was established in 1952 by the English architect and cryptographer Michael Ventris, working with the linguist John Chadwick, though acceptance of this is not yet universal. In 1962 a large palace, destroyed by fire about 1450 BC at Zákros in eastern Crete, was discovered. In 1967 the Greek archaeologist Spyridon Marinatos followed up Fouqué's explorations with excavations at modern Akrotíri on the south coast of Thera. He uncovered a whole town buried under the volcanic eruption and so preserved in wonderful detail.

EARLY AEGEAN CIVILIZATIONS

Early cultures. *Paleolithic (Old Stone Age).* Chipped stone tools made by Paleolithic hunters have been found in many parts of mainland Greece, but none are yet recorded from Crete or the other islands. As elsewhere in Europe, the latest Lower Paleolithic industries evolved into Upper Paleolithic ones with diminutive stonework. The excavations of Thomas W. Jacobsen at the Franchthi Cave on the Bay of Argos showed that boats already sailed to the island of Melos north of Crete for obsidian, a volcanic glass invaluable for early tools, by about 13,000–11,000 BC and that the cultivation of hybrid grains, the domestication of animals, and organized community tuna hunts had already begun.

Neolithic (New Stone Age). If radiocarbon dates are to be trusted, agriculture was being practiced in some parts of the Aegean area as early as the 7th millennium. The first agriculturalists in the Aegean, like those of Anatolia and Palestine, may have been ignorant of the art of making fired clay vases—traces of agricultural settlements without pottery have been identified at several places in Thessaly and at Knossos in Crete. The island of Crete appears to have been uninhabited before this time, and the first agriculturalists must have reached it by sea from western Anatolia or from somewhere more distant. Other immigrants from the east may have brought agricultural techniques and ways of life to the mainland, where they mingled with the Upper Paleolithic hunting peoples. For

Ecology of the Aegean

human habitation the Aegean is one of the most favoured regions of the Mediterranean basin. Immigrants from the coastal areas of Anatolia, Syria, or Palestine would have found the climate and ecology similar to what they had known in their homelands. The olive and vine, sources of oil and wine, the staples of the Mediterranean diet, grow in most parts of the Aegean area and may have been native there. Water, which is a problem in the present century, was probably more abundant in early times when forests were more extensive than they are today.

Agricultural communities were eventually established in every part of Greece. They made pottery by hand and ground stones to shape edged tools, axes, adzes, and chisels. Wheat, barley, oats, millet, lentils, and peas were among the crops grown, supplementing wild grapes, pears, nuts, and honey. The inhabitants continued to hunt and fish, although they also raised cattle, sheep, goats, and pigs. Arrowheads of chipped stone were used on the mainland and in the Cyclades, but none is recorded from Crete, where bone points may have served to tip arrows. Another long-range weapon was the sling, and clay sling pellets were made in Thessaly where suitable beach pebbles were not available. In Crete, clubs were armed with stone heads as in Egypt and elsewhere in the Middle East in early times. Houses with rectangular rooms are attested at Knossos in Crete, at Saliagos in the Cyclades, and at Nea Nikomedia in Macedonia. Some Aegean communities, however, may have lived in circular huts of the kind found in predynastic Egypt and in early Syria and Cyprus. By the Middle Neolithic, there existed independent walled acropolis towns with specialized industries like potteries; Sesklo is an important site several acres in extent, with nearly 30 houses, a sophisticated gate, and striking red-and-white pottery. In the Late Neolithic, walled communities with special big houses that had megarons (central halls), as at Dhimini, suggest social hierarchies and dominant chiefs.

Several Thessalian settlements were surrounded by defense walls or ditches. Copper tools—simple, flat axes and knives—were in use before the end of the Neolithic both in Crete and on the mainland.

The Bronze Age. *The Early Bronze Age (c. 3000–2200).* The transition from Neolithic to Bronze Age in the Aegean was marked by changes in pottery and other aspects of material culture. These changes may reflect the arrival in Crete and the Cyclades of new people from lands farther east bringing knowledge of metalworking with them. In Crete and the islands, the changes that inaugurated the Bronze Age were more or less contemporary with the beginning of dynastic times in Egypt. The Bronze Age in the Peloponnese appears to have begun later under the influence of settlers from the islands. The Bronze Age in central Greece and Thessaly may have begun later still. Evolved types of metal tools appear to have been current considerably before the end of the Neolithic there.

Flourishing metal-using cultures were established by the middle of the 3rd millennium in Crete, the Cycladic islands, and the southern part of the mainland. Each of these three cultures had its own distinctive characteristics; however, they had much in common, and their peoples may have spoken the same or similar non-Greek languages. Many place-names throughout the Aegean—notably ones ending in *-nt-* and *-ss-*, such as Corinth and Knossos—seem to reflect a time when a group of related languages with probable Anatolian affinities was spoken there before the introduction of Greek. A large number of words came to be adopted into Greek from this earlier language group.

Early metal tools

These Early Bronze Age peoples of the Aegean seem to have employed similar types of metal tools, including axes, adzes, and short daggers, but double axes may have been special to Crete. Tweezers were used for plucking facial hairs, and rectangular stone palettes for grinding face paints with small pestles made of attractive veined stones or *Spondylus* shell.

Lerna and other settlements on the mainland were eventually surrounded by massive walls with projecting towers, and neighbouring islands like Aigina or Syros in the Cyclades also had towered walls with trap gates. Houses with several rooms were being constructed in most parts of the Aegean by this time, and buildings at Knossos and at Vasilikí in Crete have been identified as the residences of local rulers. The so-called House of Tiles at Lerna, destroyed by fire toward the end of the period, appears to have been an important focus for the community. A massive rectangle two stories high, with a roofed balcony upstairs, the structure takes its name from the baked clay tiles found in its ruins. These small, flat tiles are thought to have come from a sloping roof and may be the earliest roof tiles known. Similar tiles were recovered from a huge circular structure of the same period at neighbouring Tiryns, of which only a section has been excavated, as it lies deep below the level of the later Mycenaean palace there. It was evidently a public building of some kind.

Cretan burial practices

Cretans in the Early Bronze Age buried their dead in communal tombs. These belonged to clans or extended families and might have remained in use for many generations. Traces of hundreds of burials have been noted in some of them. Caves and rock-shelters, as well as buildings of various kinds, were used as tombs. Circular tombs were characteristic of the Mesara region of southern Crete. They were built above ground, with low massive stone walls perhaps covered with logs and thatch or slabs. Some of the largest tombs, however, with a diameter of 40 feet (12 metres) or more inside, may have been vaulted in mud brick. Annexes with cult rooms were built in front of the entrances of some tombs, and others had chambers for offerings around the sides. When a tomb became full, a new floor was laid above the earlier burials, or parts of the tomb's annex were brought into use as burial chambers. Sometimes the remains of earlier burials were removed to separate buildings or enclosures nearby. Communal tombs at Mochlos on the north coast had rectangular compartments or rooms and flat roofs, such as those in contemporary houses. At Knossos, where the local rock

was soft, artificial caves were dug to serve as tombs. Everywhere in Crete the dead were normally trussed into a tightly contracted position, knees to chin. Sometimes the bodies were then squeezed into large storage jars or small clay chests or coffins. There was evidently much ceremonial in connection with burial, and, apart from objects of personal use such as seals, jewelry, and weapons left with the dead, vases with offerings were regularly placed inside or outside the tombs.

In contrast to the Cretans, the people of the Cyclades during the earlier part of the Bronze Age buried their dead in small graves that held a single body or sometimes a pair. The graves were often grouped in family cemeteries, which might be surrounded by a wall. The bodies were placed in them lying on one side in a loosely flexed position. Some Cycladic graves were small stone-built chambers with an entrance, although the standard type consisted of a box (cist) made with large slabs set on edge and roofed with slabs. There were platforms near the cemeteries in some cases, perhaps for musical performances, dances, or rites.

Mainland burial practices

Less is known about contemporary burials on the mainland. The graves there normally contained several bodies, which suggests that they belonged to families but not to large units, such as the clans that existed in Crete. Various types of mainland graves of this period are known, including chambers cut in the rock and stone-built tombs, such as those in the Cyclades. Circular cairns (heaps of stones), each covering several burials, on the island of Leucas in western Greece appear to go back to this time.

Pottery was still made by hand throughout the Aegean area. A useful type of vase first attested there at the beginning of the Early Bronze Age was a handled jug with a spout for pouring. Some of the earliest jugs from Crete have round bottoms and yellowish surfaces, as if they were copies of vessels made from gourds. Distinctive spouted bowls of oval shape nicknamed sauceboats were quite typical of the Early Bronze Age on the mainland and usually have a fine reddish or dark overall wash. Pottery with a similar wash and with the surface often deliberately mottled is found in Crete and is known as Vasiliki ware, after a site with a little "palace" where large amounts of it were recovered. The art of making stone vases flourished in the Cyclades from the beginning of the Bronze Age. The techniques used were simple and included boring with a hollow reed, which twirled an abrasive, either emery from Naxos or sand. The people of the Cyclades used their fine white marble not only for vases but also for remarkable figurines, mostly female but including men, some playing double pipes or seated on chairs with harps. While the majority of these figures are only a few inches high, some females are larger and a few are nearly life-size. Some have traces of painted decoration. These marble figures were often placed in graves, and groups of them have been found in sanctuaries, though whether they represented gods and goddesses is uncertain. They were exported to the mainland and Crete and may have been imitated there. Vessels of gold and silver were current in the Aegean by then, and a few have survived, including gold sauceboats of mainland type and gold and silver bowls from the islands. Gold and silver jewelry of this period, mostly from Crete, includes bracelets, necklaces, earrings, headbands, and hair ornaments of various kinds. Some of the finest of this early jewelry was found in communal tombs at Mochlos on the northern coast of Crete. The inspiration for it no doubt came from the east, and much of that from Mochlos, notably hairpins with flower heads, is reminiscent of jewelry from the royal tombs at Ur in Mesopotamia.

Protective seals

Seals came into use in the Aegean about the middle of the 4th millennium. Before the invention of locks and keys, seals were employed to stamp wet clay, which was used to secure doors or affix lids to storage jars or other containers. The design impressed by the seal might add the threat of magic to that of detection if the sealing was broken. Many seals resembling those current in Egypt and Syria have been recovered from early tombs in Crete. Some of these early Cretan seals were made of elephant tusk or hippopotamus tooth. Others were made of bone or soft, easily cut stones, such as serpentine and steatite. They were of various shapes, some of animals or birds or their heads, others cylindrical, adapted from Syrian versions of early Mesopotamian cylinder seals. They were engraved with a variety of designs, including abstract patterns and pictures of animals, notably dangerous lions and scorpions or poisonous spiders (*rogalidhas*) of a species native to Crete. Seals appear to have been in use in the Cyclades and on the mainland during this period, but very few have been recovered there. A stone cylinder seal from Amorgos resembles early Syrian ones. Most Aegean seals of this period, however, even in Crete, may have been made of perishable wood. Clay seal impressions, preserved by fires that destroyed buildings at Lerna, including the palatial House of Tiles, look as if they had been stamped by wooden seals with intricate interlocking designs.

Pictures of boats with many oars or paddles were drawn among spirals (waves?) on clay vases of this period in the Cyclades and on the mainland. The boats have a high prow often surmounted by a fish ensign, the stern being low with the keel apparently projecting beyond it. Similar vessels, though with a single mast for a square sail in addition to oars, are represented on early Cretan seals. Ships of this kind would have been capable of voyages to Syria and Egypt, whence skills and fashions were reaching Crete along with imports such as Egyptian stone vases and Syrian daggers.

End of the Early Bronze Age on the mainland (c. 2200–2000). The comparative unity of incipient civilization in the Aegean area was eventually shattered by new movements of people into the Cyclades and the southern part of the mainland. Toward the end of the 3rd millennium, many of the settlements on the mainland, such as that at Lerna, were destroyed by fire, and the houses built afterward were of a different type and more primitive. These new houses were long and narrow, only one story high, and apparently gable-roofed. The entrance was at one end, and there was often a small compartment, which might be semicircular (apsidal), at the other. The new houses were evidently built by foreign invaders settling in the places they had destroyed. Some of the previous inhabitants, however, may have survived as hewers of wood and drawers of water. A new formal dark, burnished pottery appeared, as well as a simple ware with a linear pattern on a light ground; sauceboats, however, disappeared. This pottery has many features in common with that of the succeeding Middle Bronze Age; thus there may be ethnic affinities. The site of the House of Tiles appears to have been reserved as sacred or unlucky ground, with a ring of large stones above its burnt ruins.

The Middle Bronze Age on the mainland (c. 2000–1550). The mainland was disrupted again about 2000 BC with new levels appearing at sites such as Lerna in the Argolid and Eutresis in Boeotia; there seem to be new burial habits on both coasts. Some scholars see an intrusion from the north of "Indo-Europeans," but this is a difficult, perplexing topic. Some handmade pottery may have Balkan affinities, and there is string-impressed ware at a few places that resembles in some ways the pottery of the Black Sea region. In any case, the newcomers apparently were pastoralists. Although not wealthy, they may have been one source for the appearance of the horse in Greece, an established fact before the Shaft Grave Period. Many scholars view this wave, which covered most of Greece, as representing "the coming of the Greeks"; others regard the Greek language as a rich amalgam formed within the confines of Greece and not imposed from outside. A new pottery appeared on the mainland: a class of gray burnished ware, wheel-made, with sharp angular shapes copied from those of metal vases. The polished gray surfaces of this "Minyan" ware (as it was named by Schliemann after the legendary inhabitants of Orchomenus in central Greece, where he first came upon it) look as if meant to imitate silver; later, some pieces were coloured red or yellow. After some time, "Matt-painted" pottery also appeared, again with simple linear patterns on a light ground. The traditional "long house," often apsidal, was the preferred architectural form; by the end of the period, some villages were walled.

Minyan and Matt-painted pottery

The level of cultural attainment seems low, and not much metal circulated at first. The newcomers quickly

developed connections with the islands and Crete; they imported Cretan vases, and some local vases show mainland ships. Minyan and Matt-painted pottery has been found in the nearer islands and even as far as Crete and the Anatolian coast. Burials grew from single interments to larger "family" chambers at Eleusis in Attica and on both coasts; in Messenia, in parts of the Argolid, and at Marathon there appeared a novel kind of multiple burial, with individual cists (burial chambers) or pithoi (large earthenware jars), the whole cluster being covered by a single mound. These tumulus burials, which had already appeared earlier at Leucas in the Ionian Sea, may reflect Balkan practice. In Messenia a Late Bronze Age beehive, or tholos, tomb was cut into the older mound as though that particular burial place were special. By the end of the 17th century, the newcomers had taken their full place on a newly emerging international scene and were always to be in a special relation with the Cycladic islands, Crete, and, probably, Troy. Bronze knives and gold ornaments were found with some burials, and, by the time of the Mycenae Shaft Graves in the 16th century, a luxuriant style of native goldwork had been created.

The Cyclades. On the island of Cythera (Kíthira), between western Crete and the southern tip of the Peloponnese, a colony of Cretans appears to have replaced a settlement of people from the mainland toward the end of the 3rd millennium. In the 17th or 16th century, Cretan colonies were established at Triánda in Rhodes and at Miletus on the western coast of Anatolia. Later Greek legends seem to refer to colonies from Crete, if not from Knossos, in some of the Aegean islands. Much Cretan pottery found its way to the Cyclades and was also imitated there; but, although the Cycladic people adapted some fashions and ideas from Crete, they retained their own distinctive traditions. Cycladic vases are decorated with flowers, especially lilies and saffron crocus, with swallows, wild goats, and dolphins, and with warriors and strange griffins, in a lively, splashy, and colourful style. Frescoes at Ayía Iríni (Aghia Eirene) on Ceos (Kéa) show blue birds, a town, hunting, a girl picking flowers, myrtle branches, and a copper ingot, and those at Phylakopi on Melos depict women in clothes embroidered with birds, fine textiles, flying fish, and lily blossoms. At Akrotíri on Thera, a town buried under a volcanic eruption about 1500 BC, there are in almost every house fairly well-preserved frescoes displaying wonderful, flat, brightly coloured scenes of boxers, fishermen, antelopes, birds, and blue monkeys. The two most dramatic ones are the "naval" or "miniature" frescoes from the West House, showing themes of war and peace in a seaside-and-country setting with whole towns watching elaborate ships, and the elegantly drawn set in Xeste 3, of girls and women picking saffron crocus, wearing their finest gold and rock crystal jewelry and elegant costumes; they are accompanied by blue monkeys. The Theran paintings are the best surviving Aegean documents for clothing, architecture, ships, armament, and daily life.

The Shaft Grave Period on the mainland (c. 1600–1450). There are links between the Thera paintings and such items as earrings, necklaces, and metal vessels found in the royal Shaft Graves at Mycenae. Thera itself, however, had few valuables like metal; apparently the inhabitants had taken prized objects away. The Shaft Graves, in contrast, were packed with gold, silver, and bronze—almost nomadic in the obvious preference for portable gold and weapons. Two groups of Shaft Graves were discovered at Mycenae in different parts of a large cemetery area. The burials in them seem to have ranged over a period of 150 years, from shortly before about 1600 to the middle of the 15th century. Each group was eventually surrounded by a circular enclosure wall. The circle designated B, with the earliest burials, lay outside the limits of the later Bronze Age defenses, but the other circle, A, enclosing the richest burials in six large graves, was deliberately incorporated within them. The wealthy burials belonged to leading, if not royal, families of the place that would eventually supplant Knossos as the chief centre of the Aegean. Schliemann excavated the graves of Circle A in 1876, but it was not until 1951 that Circle B was noted. These graves are capacious shafts cut in the rock, often with pebble floors and slab roofs. They were used for multiple burials over a course of at least several years, and the remains, including beef bones and oyster shells, give evidence of well-developed funeral rites. Both men and women were buried in the graves, many of which contained several bodies. After the bodies were placed in the graves, the stone-walled burial chambers were roofed with timbers, and the shafts above were filled again. Sometimes the remains of earlier burials seem to have been pushed aside to make room for later ones, but, if so, the shafts must have been laboriously reopened to admit new burials. Large stone slabs with carvings in flat relief had been set above some of the graves. The carvings include spiral designs and pictures of the dead riding in their chariots to war or to the hunt. They have vivid battle imagery—three stallions rearing, spears ground under chariot wheels, and a man falling headfirst from a chariot. In one case, the scene of a warrior driving a chariot over a fallen enemy encased in a shield seems to be reinforced by a scene just below it, a lion chasing a deer. This visual simile may be analogous to lion similes in Homeric epic. According to another interpretation, the dead were taking part in their own funeral games with chariot races as described in Homer. These tombstones provide the earliest evidence for chariots on the mainland.

Two groups of Shaft Graves

Earliest evidence of chariots on the mainland

A fantastic array of gold and silver cups, jewelry, and dress ornaments had been placed with the dead, especially with those in the graves of Circle A. Golden diadems and elaborate hairpins decked the heads of women. Beads in necklaces were of amethyst, probably from Egypt, and amber, from the Baltic. The men were buried with supplies of bronze weapons, including great slashing knives and spearheads and two kinds of rapier-like swords, a mainland version and a Cretan version. Several swords are ornamented with gold-plated hilts and pommels of polished stone, ivory, or gold; some have gold predators at the hilt gripping the blades in their mouths. The blades may be ornamented with running horses, flying griffins, shields in the shape of a figure eight, or even lilies running down from hilt to tip. The tremendous influence from Crete on these graves is visible in the metal cups, faience "sacral knots" (*i.e.*, representations of a Cretan ritual object in the shape of a scarf with a looped knot and fringed ends), dolphin-appliquéd ostrich eggs, conch shells associated with ritual summoning, gold triple shrine facades, images of bulls with double axes between their horns, and imported pottery painted with plants. Beside them is an equal wealth of local art such as formal gold cups, gold worked in breathless surface patterns of lions, bulls, and plants, and dozens of lions twisted as ornament. There probably is a local iconography in the gold seals of duels, lion combat, chariot hunting, and a wounded lion trying to pull an arrow from his shoulder. Traveling artists may account for some of the similarities to Cycladic and Cretan art, but local armourers may also have wrought local metal into drinking cups. Covering the faces of some of the men were gold portrait masks showing them with beards and mustaches. In this they are like an amethyst "portrait" gem in Circle B of a bearded mature man. (Later studies of faces also seem to reserve the beard and mustache for important or powerful elders, although fashions change; servants and soldiers are normally beardless). Women's costume cannot be known from the remains, but it may have had the same range of tunic, apron, and veil as in the Theran paintings, and the jewelry is impressive.

Some bronze dagger blades were inlaid with remarkable pictures or designs in other metals, chiefly gold and silver and electrum (a mixture of gold and silver) in various shades. Black niello was used as a background for these pictures or to heighten the incised detail. The most famous of the Shaft Grave daggers shows men armed with bows, spears, and great body shields, hunting a pride of lions; another has catlike animals chasing wild fowl among papyrus flowers beside a silver stream. This technique of "painting in metals" appears to have originated in Syria, although the workmanship and style of the pictures on the Mycenae daggers look novel. Whereas other daggers and some metal cups with inlaid designs of this kind have been found on the mainland, none has yet been recovered

in Crete. But many of the treasures from the Shaft Graves are imports from Crete.

The Shaft Graves had so many metal vases, including huge bronze cauldrons (one marked with Linear signs), that clay vases were not much needed. Yet, the contemporary chamber tombs at Mycenae and many other sites have wonderful pottery that is both imported from Crete and made with local taste with spirals, ferns, and double axes. In this development one can observe the formation of a new Mycenaean Greek culture, as it assimilated styles from Crete and yet insisted on more traditional local habits. It is this tentative fusion of two cultural "languages of art," already in touch for two or three centuries, that gave a special impetus to the new Mycenaean world, rendering it flexible, receptive, and adventurous. The pottery, superior in technique, colour, and design, was attractive to other cultures and widely used as commercial containers for oils. Because it has been found in almost all coastal districts from Syria to Sardinia, it is a real aid to dating.

Tholos tombs

Along with the rich chamber tombs at Mycenae, certain families, perhaps princely, began building tholos, or beehive, tombs as early as the Shaft Grave Period, perhaps first in Messenia in the 16th century and then in many places in Greece by the middle of the 15th century. The tholos tomb has three parts: a narrow entranceway, or dromos, often lined with fieldstones and later with cut stones; a deep doorway, or *stomion*, covered over with one to three lintel blocks; and a circular chamber with a high vaulted or corbeled roof, the thalamos. When the facades are finely dressed with cut stones or recessed vertical panels, one may think of a Cretan connection; indeed, one of the tholos tombs at Peristeria has two Cretan "masons' marks," a branch and a double ax, cut into the facade to the left of the doorway. The influence of Crete on the southwest Peloponnese is marked. Perhaps a traditional memory of this connection is preserved in the Homeric Hymn to Apollo, which tells of the god kidnapping the crew of a ship trading from Knossos to Pylos to serve his new sanctuary at Delphi. Excavations at Delphi yielded a snout of a marble lion rhyton (libation vessel), matched best by a complete example at Knossos. The tholos tomb is always covered by a mound of earth, often kept in place by a peripheral stone ring, or *krepis*. Some tholoi were built on the surface of the land, but most were built in a deep pit excavated into the slope of a hillside. The stones that were overlapped in rings to form the vault in the corbeled system were laid with a narrower face inside, which locked each ring in place. The lintel blocks, often huge in size and weight, were dragged across the hill and dropped onto the corbel rings at the proper height; either a single huge block or two or three slabs next to each other provided the needed depth. Various systems were used to ease the weight on the lintel, such as narrow stone bars or an open relieving triangle sealed by a thin-cut screening stone. The whole vault was sealed with a keystone.

Most tholos tombs have collapsed, often when the lintel cracked and gave way, and their contents have largely been looted. Occasionally the robbers overlooked a pit sunk in the floor, like the rich burial at Vapheio near Sparta; sometimes a whole tomb survived unplundered, like the one at Dendra near Mycenae or that at Rutsi-Myrsinochorion in Messenia. Of the nine tholos tombs at Mycenae, two, the Treasury of Atreus and the Tomb of Clytemnestra, have splendidly dressed facades with engaged half columns in two tiers and coloured exotic stones; they may have been built early in the 14th century, although arguments are made for a 13th-century construction. The elaborate design of the facade may have been imitated from the impressive north facade of the inner court at the Cretan palace of Phaistos. The imagery might imply a continuing presence of the dead kings inside the tomb. Such tombs sometimes mark the gate or main road to a town, as classical tombs did, as though they were "ancestral watchers" or guardians. Tholos tombs were built from the 15th to the 13th century and imply a hierarchical command of labour, of the kind the palace exerted later, according to the Linear B documents. Possibly the capstone was not put in place until the dynast died. These structures could not be built quickly but were prepared with foresight.

While stone-built tholos tombs became the standard resting places for kings and princes in all parts of the mainland to which the Mycenaean civilization penetrated, the mass of the population changed from a custom of burial in single graves, whether in mounds or cemeteries or inside settlements, to the use of family vaults. In some regions, such as Messenia and the frontier area of Thessaly, families built small tholos tombs for themselves. The most common type of Mycenaean family tomb, however, was a rock-cut chamber with a dromos leading down to the entrance. The entrance was blocked with stones and the passage filled with earth after each burial. The rock-cut tomb may have been developed in Messenia during the 16th century under Cretan influence, like the tholos tomb. In the Knossos region of Crete, rock-cut tombs had been in use for communal burials for many centuries before this. Whatever its origin, the idea of family burial in rock-cut tombs soon spread to Mycenae and other parts of the mainland. Some rock-cut tombs in Messenia and elsewhere were carved in the shape of the beehive vaults of tholos tombs. A few large rock-cut tombs, including some of this shape, were used for royal or princely burials.

Period of the Early Palaces in Crete (c. 2000–1700). Crete does not seem to have been affected by the movements of people into the Cyclades and the mainland at the end of the 3rd millennium, but important changes were taking place there. Great palaces of a distinctive type built around large rectangular open courts seem to have been constructed within a comparatively short time at the leading centres of Knossos, Phaistos, and Mallia. The art of writing is first attested for certain in Crete at the beginning of this Palatial Period. These developments in Crete appear to have been the result of local evolution.

Earliest writing in Crete

Crete advanced rapidly along the path of civilization during the period of the Early Palaces, while the mainland relapsed into comparative agricultural stagnation. The art of seal engraving made great strides in Crete. Hard stones, such as jasper and rock crystal, began to be employed for some of the finer seals. A new and much-favoured shape, which may have been adopted from Anatolia, was the signet with a stalk. Anatolian seals found their way to Crete, and impressions of them have been identified in a great deposit of clay sealings from the early palace at Phaistos. Cretan seal designs now included elegant abstract patterns of spirals and concentric circles neatly made with the drill as well as lifelike pictures of animals, birds, and insects, together with mythical beasts such as sphinxes and griffins adapted from Egyptian or Oriental models. Attractive hard stones, such as gabbro, were used by the Cretan vase makers, although they still used the softer chlorites and serpentines. Some of the fine stone vases from communal tombs in the Mesara region and at Mochlos may date from this period, rather than earlier, in the light of discoveries since 1950 in the early palace at Phaistos.

The fast potter's wheel began to come into use in Crete about the same time as in the Cyclades and on the mainland. Meanwhile, a revolution in the style of Cretan pottery was taking place. During the Early Bronze Age most of the finer vases everywhere in the Aegean area had been decorated with designs in dark, rather shiny paint—shades of red, brown, and black—on a light surface. Toward the end of that period in Crete, however, there was a change to a "light-on-dark" style of decoration; the vases were given an overall wash of the shiny paint previously used for decoration, and designs were applied to this dark surface in white. This new light-on-dark fashion was also adopted, to some extent, in the Cyclades and on the mainland, but in Crete it was developed much further, and, from the beginning of the Palatial Period, decoration in white was regularly supplemented with red to create a striking polychrome effect. This kind of pottery, which flourished in Crete throughout the time of the first palaces and later (*c.* 2200 to 1600), is known as Kamáres ware from a sacred cave of that name on Mount Ida, where vases with fine polychrome decoration were recovered at the end of the 19th century. Most of the smaller vases in Crete, notably the drinking cups, now copy metal ones in their shapes and often in their molded or impressed decoration, and the exquisite "eggshell" ware, made in the

"Eggshell" ware

workshops of the great palaces, with walls as thin as those of metal vases and shiny black surfaces adorned with abstract flowerlike designs in a combination of white, red, and orange, is among the finest pottery ever produced in Greek lands. The imitations in clay suggest that vessels of precious metal—gold and silver—were in general use in the palaces of Crete by this time. A silver, two-handled goblet of this period was recovered from a tomb at Gourniá in eastern Crete. Silver occurs in the Cyclades, and it was being mined during the Bronze Age near Laurium in Attica on the mainland.

There were many contacts between Crete and the rest of the Levant during this period. Scarabs and stone vessels from Egypt reached Crete and were imitated there. Cretan Kamáres ware was exported to Cyprus, Syria, and Egypt, where it has been found in tombs and on town sites. Letters recovered from the ruins of the city of Mari on the Euphrates, destroyed by Hammurabi about 1760, refer to objects of Cretan workmanship. It seems that Cretan metalworkers were already preeminent in the civilized world of the time. The daggers they made were of types ultimately derived from Syria, but they were exported to Cyprus in exchange, perhaps, for copper, although supplies existed in Crete. Westward, they may have reached Italy, where native copper daggers are of Cretan shapes and flint imitations of them seem to have been made. It was during this period that tin-bronze began to come into more general use in the Aegean, replacing copper or bronze made by adding arsenic, a process which was effective but dangerous for the craftsman who undertook it. Tin may have reached the Aegean first from Iran through Syria, although Etruria on the western coast of Italy was another possible source.

Burial in Crete was still normally in communal tombs, and many of the Early Bronze Age ones continued in use, but cemeteries of burials in storage jars are also in evidence at this time. No royal tombs of this period have been identified, however, and kings and queens may have been laid to rest, like their subjects, in the tombs of their clans or possibly even buried ceremonially at sea. A large rectangular building with many rooms or compartments in the cemetery area just outside the city at Mallia might have been the tomb of the royal clan there. The local inhabitants plundered it during the 19th century, and its modern name—Chrysolakkos ("Gold Hole")—suggests what they found. A gold cup and jewelry, including elaborate earrings and pendants, acquired by the British Museum in 1892 and allegedly from a Mycenaean tomb on the island of Aegina near Athens have been thought to be plunder from Chrysolakkos, although recent excavations on Aegina have indicated a wealthy and warlike community that could equally have produced these jewels. They are marked by an unusual style: one earring has a two-headed snake surrounding a pair of leashed hounds over squatting monkeys, with owls and discs hanging on soldered chains. The collection may have been made during the 17th century, after the destruction of the older palaces. French excavations there in the 1920s led to the recovery of similar jewelry, notably a gold dress-pin with flower head and a pendant in the form of a pair of bees (or wasps) facing each other over a disc, which may be meant for a honey cake. This pendant shows that the Cretan jewelers were masters of the art of hard soldering and could use it to fix wire (filigree) or minute globules of gold (granulation) to a background.

Aegina treasure

Life in the Cyclades seems to have continued much as it had in the Early Bronze Age. Yet, apart from signs scratched or painted on pottery from Phylakopi in Melos, there is little evidence of acquaintance with writing or the use of seals. Some time after the beginning of the period of the Early Palaces in Crete, Phylakopi was defended by a massive wall. Cretan Kamáres ware was exported to the islands of Melos, Ceos, and Aegina and to Lerna and a few other coastal sites on the mainland, and mainland Minyan ware found its way to the islands and to Crete. The trade may partly reflect the trade in Melian obsidian, which may still have been in demand for cheap knives and razors, although metal ones were already in use in the Aegean area from the Early Bronze Age onward. Chamber tombs cut in the rock at Phylakopi appear to go back to this period, but burial in slab-lined cists continued elsewhere in the islands. At some point the fortified settlement at Khalandriani on Syrus was destroyed by fire and abandoned, but Aegina, Ceos, and other fortified island towns flourished.

Period of the Late Palaces in Crete (c. 1700–1450). Various disasters occurred in Crete about the turn of the 18th and 17th centuries BC. The palaces at Knossos and Mallia were damaged, while that at Phaistos and a building that may have been the residence of a local ruler in a large settlement at Monastiráki west of Mount Ida were destroyed by fire. The palace at Phaistos had been so violently burned that an enormous layer of almost impenetrable vitrified mud brick formed an underpinning for the new palace built on top of it; it is a vivid testimony to massive destruction. What caused these destructions is uncertain. Accident, internal warfare, or foreign invasion are among possible agents. The damage at Knossos might have been caused by one of the many earthquakes that afflict the area. It has been suggested that Crete was first conquered by Greeks during this period or by people from Anatolia speaking another Indo-European language called Luwian and related to Hittite. There is, however, no strong evidence for an invasion of Crete at this time.

Destruction of the palaces

The two or three centuries following these disasters were indeed the most flourishing of the Aegean Bronze Age, during which Cretan civilization reached its zenith. The palaces at Knossos, Phaistos, and Mallia were restored with greater splendour than before.

From the dimensions of the new and entirely rebuilt palace at Phaistos, it has been possible to calculate the unit of length used by the Cretan architects: a foot only a fraction shorter than the standard English foot. In plan, the later palaces were basically the same as the earlier ones, with agglomerations of rooms clustering around long, rectangular central courts oriented roughly from north to south either for ritual or for catching the best of the winter sun. Many parts of these palaces were two or three stories high. A section on the eastern side of that at Knossos, built into a cutting in a steep slope below the level of the central court and housing the royal living quarters, may have had five stories. Large areas of the palaces, especially at Knossos, were possibly reserved for cult. It is difficult to explain otherwise the beautiful ceremonial steps at Phaistos leading up to a blank wall; although there is no entrance, a personage could make a sudden appearance from the side and speak or show something to an assembly in the open space in front. The palaces often had a conjunction of

De Antonis



Vats and cists for storage in the west magazine of the second palace at Knossos, Crete.

grand facades and storage quarters, perhaps for the first fruits of the harvest to be blessed in passing.

Wide, paved squares flanked the palaces, and around them spread extensive towns, which by this time if not earlier seem to have been unwalled. Unfortunately, a complete town around a palace has never yet been excavated, and the comparative wealth or population is not known. Cobbled streets with raised central paths of smooth squared blocks for the convenience of pedestrians ran through the towns. Surface water was carried away by covered drains, and skillfully jointed clay water pipes were found in the palace at Knossos.

Gourniá

The only settlement of this period that has been entirely excavated is a small town at Gourniá in eastern Crete. This was built on the slopes of a ridge overlooking the sea, on top of which stood a little "palace" with a small open court in the centre and a public square beside it on the sheltered landward side. Down the ridge from the palace toward the sea was a small shrine facing the end of a path that led to it from the main street. Even in a small town such as that at Gourniá, many of the houses were evidently two stories high, and houses with three stories are depicted on faience inlays from the palace at Knossos assignable to the 17th century BC.

Palaikastro in eastern Crete is another important town with blocks of houses marked by coloured stone foundations, narrow streets with drains, and pottery of exceptional quality. Another town of great potential interest is Arkhanes near Knossos, where palace facades, early tholos tombs and later shaft-grave burials, and shrines have been discovered scattered through the countryside. Pyrgos, a controlling villa, and Kommos, a commercial town with fine architecture, roads, and ship sheds, also are indicative of power and wealth; the road and watchtower system is beginning to be better known.

In palaces as well as houses, the lower parts of walls were still normally built of rough fieldstones held together with mud, the upper stories being continued in mud brick. Carefully squared and fitted blocks of limestone, however, were employed for some important facades. Now, as earlier, walls were often tied together with a framework of timbers set vertically and horizontally and joined by crossbeams running through them. There was also an extensive use of timber for columns and pillars and for the rafters supporting upper floors and roofs, which, it seems, were usually flat. Pictures of wooden columns show them with a characteristic downward taper, which may reflect an original custom of placing tree trunks upside down. The lower parts of the walls inside the palaces and great houses were often clothed with large slabs of attractively veined gypsum, a soft crystalline stone that outcrops in the region of Knossos and Phaistos. Gypsum was also much employed for pavements, but a hard lime plaster was more commonly used for coating walls and floors. Plastered walls were decorated with brightly coloured pictures, which may be an innovation of this period, since they are not yet attested for certain earlier in Crete.

Minoan frescoes

These pictures are described as frescoes because they were normally painted while the plaster was still damp. Lines impressed with string in the wet plaster helped to guide the artists. White, red-brown, or blue were usually chosen as a background, while yellow and black were among the other basic colours used. Many of these pictures, especially those from the palace at Knossos, were concerned with religion; they show elaborately dressed goddesses, together with sacred dances and ceremonies, such as bull leaping, which appears to have had a religious or magical basis. Yet scenes such as a frieze of partridges and hoopoes adorning a room in what seems to have been an inn for strangers opposite the palace at Knossos look entirely secular. Monkeys, imported from Egypt, are depicted more than once, along with native wild goats and extraordinarily lifelike flowers—rose, ivy, saffron crocus, lily, and papyrus—but often imaginary hybrids. Some frescoes may represent permanent magic gardens. The pictures ranged in scale from those with life-size figures, which might occupy most of the wall surface, to panels and friezes, including a class of miniatures with figures of men and women two to three inches (five to seven centimetres) high. Parts of some wall pictures at Knossos were in relief, and plaster reliefs of this kind are occasionally found elsewhere in Crete. Floors and ceilings might also carry painted decoration.

Their wall paintings were probably the finest achievements of the Cretan artists, but only battered or fire-discoloured fragments of these have survived. The minor arts are better represented in the archaeological record. Now, if not earlier, hard rock crystal began to be used for making vases and seals, together with the volcanic glass, obsidian. A variety flecked with spots of white pumice, from Yialí (Glass Island), near Cos, was favoured for vases. Other fine stones imported for vase manufacture were Egyptian alabaster (calcite) and green and red marbles (antico rosso and lapis lacedaemonius) from the southern Peloponnese. Antique stone vases from Egypt might be adapted to local tastes by the addition of spouts and handles. Vessels with narrow necks were carved in two pieces that were afterward joined together, an example being a crystal libation vase from Zákros with the handle formed of crystal beads threaded on copper wire. A number of cult vases are carved with pictures in relief, including an octopus, a mountain shrine with birds perched on horns of consecration, altars in an enclosed courtyard, and wild goats and, on other vases, youths engaged in ritual competition, a ritual dance of some kind, and games, such as bull leaping, wrestling, and boxing, which apparently had magical or religious connotations. Soft stones, such as chlorite or serpentine, were used for making these vases, the surfaces of which were often coated with gold leaf, to judge from the scraps that have survived. This economical system of gilding was sometimes applied to seal stones, although solid gold and silver seals also occur. A class of gold signet rings has oval bezels engraved with ritual scenes that may be from the story of a goddess and her consort and include scenes of worship at an altar or a tree, with a shield or sacral knots as attributes, or dancing. Seals of other shapes, in a wide range of attractive stones, display a variety of designs, including animals, such as lions, bulls, and wild boars. Sometimes a bull is being attacked by a lion, or a wild goat is escaping or standing at bay before a hound. Birds, fish, and butterflies also figure on these seals, and most of the designs appear to be entirely secular in character. A class of gems crudely engraved with pictures of jars and leafy branches may have been rain charms, however.

Rings and seals

There is little evidence for Bronze Age sculpture in Crete, apart from a few small stone heads that may have come



Faience statuette of a so-called snake goddess from the temple depository of Knossos, c. 1600 BC. In the Archaeological Museum, Iraklion, Greece.

Ministry of Culture, Archaeological Receipts Fund, Greece

from statues with wooden bodies or a pair of clay feet perhaps supporting a dressed armature. Some bronze curls from the palace at Knossos appear to have adorned the head of a more than life-size wooden statue of a goddess. Figurines cast in solid bronze, though sometimes marred by casting defects, are often of great beauty. They mostly represent worshipers, both men and women, and were placed as votives in sanctuaries. Statuettes of bull leapers and perhaps of gods and goddesses were made of imported ivory in several pieces cunningly joined together by pins and dowels. Faience manufacture was presumably learned from Egypt. Exquisite faience plaques of animals, along with statuettes of goddesses or priestesses and small vases of the same material, appear to be products of the palace workshops at Knossos for shrine or ritual display.

The Late Palace Period seems to have been rich in metals. Although few gold and silver vessels have survived in Crete, many fine vessels in the Mycenae Shaft Graves may have been made by Cretan skilled workers. Even cooking vessels were now being made of copper or bronze, including huge cauldrons in which a sheep or goat could be boiled whole. Among a variety of serviceable bronze tools were axes, adzes, and double-bladed axes such as those of earlier times. The sockets of these were improved toward the end of the period from a circular to an oval shape, which prevented twisting of the haft. New tools current by then included long bronze chisels and immense saws capable of slicing the gypsum required for paving and wall veneer, as well as for cutting timber. Helmets of copper or bronze are depicted on faience inlays from Knossos and on stone relief vases, but plate armour is attested only from the end of the 15th century. For defense, the Cretans of this time, like their Mycenaean and Cycladic contemporaries, appear to have relied on huge rectangular or eight-shaped shields of bull's hide. (Homer's description of the shield of Ajax as being "like a tower" preserves a memory of body shields of this kind.) Weapons included spears and daggers, as well as rapiers with long slender blades and short tangs for affixing wooden hilts. Massive pommels of attractive stones, such as rock crystal, or of gold-plated wood or ivory helped to balance the blades. Toward the end of the period, swords are found with strong, flanged hilts and short blades adapted for cutting as well as thrusting strokes. A remarkable set of weapons, often inlaid, enriched with gold, ivory, and designs, was created at Knossos at one or more brilliant sword workshops (which vanished after about 1400).

Signs scratched or painted on clay vases, not only in Crete but on the mainland and in the islands, from about the middle of the 3rd millennium onward may reflect acquaintance with writing among the peoples of the Aegean area. The first positive evidence for the use of writing in the Aegean, however, is found in Crete at the beginning of the Palatial Period—about 2000 or somewhat later. This earliest Cretan writing is known as pictographic or hieroglyphic because its signs are pictures of animals or things; the system appears to be of Cretan origin, even if it was inspired by Egypt or Syria. During the period of the Early Palaces and while the Cretan hieroglyphic script was still in use, a simplified linear script was being scratched on clay tablets at Phaistos. A more evolved script with linear signs of this kind is attested in various parts of Crete and was known in the Cyclades during the Late Palace Period. It is known as Linear A to distinguish it from the variety of script (Linear B) current both in Crete and on the mainland from the end of the 15th century (see below). Most of what has survived of Aegean Bronze Age writing is on clay tablets of the kind used in Syria and Mesopotamia in early times. Ink was, however, used to write Linear A inscriptions around the insides of two clay cups from Knossos, and the bulk of what was written in the Aegean during the Bronze Age may have been in ink on some kind of paper made from papyrus, as in Egypt, or from palm leaves, as later Greek tradition hints. The two standard forms of tablets are the long narrow "palm leaf" for short transactions and the tall rectangular "page," which often is a summary or inclusive list. The Knossos tablets supply records of transactions involving personnel, cattle, sheep, goats, oils and spices, wool and textiles, weapons (including arrows, swords, and issues of chariots with armour), stored treasures, and religious offerings. They seem to reflect a period when the former palaces of the several districts were no longer standing, or powerful, but when the surrounding lands still produced agricultural goods that were taxed or tithed at Knossos.

Linear A and Linear B

THE DECLINE OF THE EARLY AEGEAN CIVILIZATIONS

The eruption of Thera (*c.* 1500) and the conquest of Crete (*c.* 1450). Cretan civilization reached its highest peak between about 1600 and the later 15th century. An important change of fashion that began about 1600 in Crete was the abandonment of the "light-on-dark" style of vase decoration of Kamáres tradition in favour of a return to "dark-on-light." The new-style Cretan pottery, with attractive designs of spirals, grasses, ferns, and flowers in shiny black or brown paint, was soon to inspire the development of Mycenaean pottery on the mainland. This flourishing period in Crete, however, ended in a series of disasters. About 1500 the volcano on the island of Thera, long, it seems, quiescent, erupted to bury the settlements there under many feet of pumice and ash. The story of Atlantis, if Plato did not invent it, may reflect some Egyptian record of this eruption, one of the most stupendous of historical times. Knossos was shattered by a succession of earthquakes that preceded or accompanied the eruption, while great waves resulting from it appear to have damaged settlements along the northern coast of Crete. Ash identified as coming from the eruption has been found in coastal sites as far away as Israel and Sardis in Anatolia. The wind may have been blowing from the south or west. Later Greek traditions, such as the story of Deucalion's flood, may enshrine a memory of similar waves that swept the coasts of the mainland at this time. Some Cretan settlements might have been wrecked by the blast from the eruption, although Thera lies about 70 miles (110 kilometres) away from Crete. Whatever the damage caused, it appears to have been soon repaired and not to have disrupted the course of local culture. Damages to pastures and livestock were apparently minimal. Similarly, in the Cyclades there are few signs of any gap in occupation as a result of the eruption. The settlements on Thera, however, lay buried deep in pumice. The largest of these, at Akrotíri, opened by excavations since 1967, offers a unique picture of a Bronze Age town. The walls of its houses stand in places two stories high, with paintings miraculously preserved on them, and the floors with storage jars and other objects are as they were left when the inhabitants escaped from the eruption or from the earthquake that is thought to have preceded it. The wonderful preservation of delicate frescoes and of foodstuffs, from snails to olives to grain, makes Thera a tantalizing closed deposit of Aegean life. Many houses have flagstone floors upstairs, with columns supporting the roof, and rooms with multiple windows. Below there are storage bays filled with jars, looms, medicine chests, and grain and oil stores. There is delightful local pottery with swallows, dolphins, wild goats, and caper and saffron plants. The wall paintings have a garden quality and may often have religious associations or celebrate festivals and seasons, city or country life, and sea voyaging. Only a small part of the town has been excavated. The work has been slowed by the engineering problems of keeping the two- or three-story houses from collapsing and crushing the painted walls and delicate contents sealed from damage for centuries.

After the eruption, Crete appears to have enjoyed comparative prosperity for a time, while the influence of Cretan civilization continued to spread on the mainland. Alongside vases with plant and flower designs, the Cretan potters began to decorate others in an attractive marine style, with octopuses and other sea creatures. The marine style may have originated at Knossos, but vases with this type of decoration, many of them of a ritual character, were exported all over Crete, as well as to the Cyclades and the mainland. About the middle of the 15th century, however, a generation or so after the eruption of Thera, most of the important sites in central and southern Crete were destroyed by fire. Destruction was not confined to

The marine style

palaces and towns but extended to country houses, farms, and rural shrines. Many settlements were never inhabited again, such as that at Mochlos, where excavators found the remains of numbers of people who had perished in the destruction. The site of the destroyed town at Gourniá was eventually occupied by a scatter of houses, but the palace there was not rebuilt. The large palaces at Mallia and Zákros were also destroyed by fire and afterward abandoned.

A new social order. The fact that palaces and country houses, centres of landed estates, were not rebuilt suggests a total overthrow of the existing social order. A number of magnificent stone ritual vases and bronze tools have been recovered from the ruins of the palace at Zákros in excavations since 1962, but virtually no gold or silver objects were found. Indeed, it looks as if the palaces and houses everywhere in Crete had been ransacked before they were destroyed. Of the four great palaces, only that at Knossos may have escaped serious damage at this time, but parts of the city there were wasted by fire. In the early days of Cretan exploration, it was taken for granted that such destruction was the result of war. Since the 1930s, however, it has been suggested that it was in some way caused by the eruption of Thera. The eruption, however, began and ended a generation or more before this horizon of destruction, while evidence of conquerors in Crete immediately after it has been found. The destruction appears to have been their work.

The conquerors evidently came from the mainland and made their capital at Knossos, but they seem to have established another centre of power at Phaistos, and legend hints at a third centre at Kydonia (modern Khaniá) in western Crete. Vases of shapes already popular on the mainland, such as drinking cups with tall stems, became fashionable at Knossos after the conquest and eventually spread to other parts of the island. A rather stiff, formal "Palace Style" of vase decoration, using motifs derived from the earlier plant and marine styles, may reflect an adaptation of Cretan fashions to mainland tastes. The old clan tombs went out of use in the Knossos region and were replaced by rock-cut tombs. Some of these contain the burials of warriors and their families, accompanied by rich assortments of weapons and jewelry, resembling the military equipment of the Mycenae Shaft Graves and the mainland tholos tombs. There was a cemetery of similar rock-cut tombs, with richly furnished burials (the Tombe dei Nobili), at Phaistos. Tholos tombs sunk in the ground and covered by mounds appear to have been introduced to Crete from the mainland now. One, on the Kefāla ridge north of the palace at Knossos, may have been built soon after the conquest. It has masons' marks like the one at Peristéria in Messenia.

The Linear B texts. Insight into the social order on Crete after the conquest can be gleaned from the Linear B tablets found at Knossos, where Linear B had replaced Linear A by the 14th century BC. The decipherment in 1952 of the Linear B tablets as Greek by Michael Ventris, working with John Chadwick (see above), has been widely accepted. Still, there are some skeptics who reject it; and, while most of these believe that the language of the tablets will prove to be Greek when (in their view) it is correctly deciphered, a minority think it will not. At the same time, among philologists who do accept the Ventris decipherment, there are a few who regard the language of the tablets as a form of Greek with little or no relation to the Greek of later times, which, in this minority view, was introduced into the Mycenaean world by new peoples of Greek speech at the end of the Bronze Age. The tablets have many personal names and place-names but very little connected descriptive Greek, making them hard to read; nonetheless, they are of enormous potential value.

The Knossos records

Knossos seems to have been the only Cretan centre with a genuine archive recording income, palace issues of expensive equipment like chariots and bronze corselets, and outlays on gifts to the gods (some of whom were Greek, some traditional Cretan), mainly in the form of donations of oil and cloth. The Knossos records are valuable in allowing reconstruction of farming practices, the wool industry, and military defense, as well as in providing lists of personnel and places scattered around the countryside of Crete that owed or brought sheep and produce to the palace. The bureaucratic apparatus seems to have been well organized and extensive, whereas the rest of Crete in the 14th and 13th centuries, though rebuilt after the earlier disasters and the abandonment of the 15th century and prosperous, does not give evidence of the same degree of control and record keeping. How far the countryside was subservient to Knossos is not known.

The fusion of cultures on Crete. The last decades of the 15th century and the first part of the 14th century saw a wonderful fusion of Cretan and mainland skills; the fabric and firing of pottery improved, and there were formidable and often elegant and richly ornamented bronze weapons in the warrior graves at Knossos and Phaistos. Sword hilts were sometimes sheathed in gold, and their pommels made of fine stones and ivory. Warriors were armed with large thrusting spears as well as throwing spears, or javelins; boar spears were used for hunting. Shields were painted on walls as well as hung there, and there were new bronze arrowheads and conical helmets with a plume knob and cheekpieces. The old mainland helmet plated with tusk plaques of the wild boar reached Crete, too. Bronze armour was issued by the palace, with chariots and pairs of horses; the transverse strokes on the cuirass ideogram on the tablets suggest a link with the transverse bronze bands of the armour suit from Dendra near Mycenae from about 1400 BC. There seems to have been a disciplined set of chariot squads, perhaps often composed of men from abroad, patrolling the island. It may be that a memory of this energetic military epoch is preserved in the *Iliad,* in the passages describing the exploits of the Cretan princes Idomeneus and Meriones. Gold rings from this period have a rich religious iconography of shrines and worshipers, and there are fine sealstones. About 1400 the fused customs or beliefs of the Cretan and Mycenaean worlds appear on the painted limestone sarcophagus from Ayía Triadha, a small palace near Phaistos; depicted on one side are libations on the left and, probably, offerings to a dead man on the right. The other side shows a bull being sacrificed while a man is playing the double pipe. Pairs of women in chariots drawn by wild goats or griffins are represented on the ends. Historically the sarcophagus occupies a dividing line: Knossos was burned again after 1400.

There seems to be a marked difference in economic power or aesthetic achievement between the earlier (1600–1400) and the later (1400–1200) periods of Minoan, Cycladic, and Mycenaean cultures. Frescoes are duller—more repetitive, coarser in outline, and muddier in colour. The grand swords are no longer made, and rings and seals are simpler. Metal becomes rarer, and blue glass tends to replace lapis lazuli from the east. It is uncertain whether this decline in art reflects a change of governmental forms, a restriction on trade with Egypt and the east, a decrease of creative energy leading to an unimaginative reproduction of traditional patterns in familiar materials, or the loss of palace-controlled Cretan workshops that had supplied the earlier fine standards for the developing Greek world.

The mainland. While there are many signs of mainland influence in Crete in the period after about 1450, the conquest may have helped to spread Cretan fashions and techniques on the mainland through the medium of captive artisans sold as slaves. The earliest wall paintings on the mainland appear to date from this time and are thoroughly Cretan in style. The Mycenaean civilization of the mainland nevertheless remained very different from that of Crete. Mycenaean pottery is distinguishable from Cretan, and religious customs, such as worship in caves or hilltop sanctuaries, which continued in Crete, do not appear to have taken root on the mainland. The sphere of architecture, however, is continually impressive, as it had been in the older phase of tholos tombs.

Mainland palaces

The standard mainland palace of this period, although built with Cretan techniques, differed from the traditional Cretan palace centred around a large, rectangular court. The focal point of a mainland palace, such as that at Pylos (Pilos) in Messenia, was a great hall with the roof supported on four pillars and a vast circular central hearth. The hall was entered through an anteroom with a

columned porch beyond it. This complex appears to be an adaptation of the type of longhouse found on the mainland since the end of the 5th millennium. The mainland palaces were painted in a manner derived from Cretan models, with large processional scenes and smaller scenes of men hunting boars or stags, of chariots, duels, and numerous battles; there are heraldic hounds, griffins, lions, sphinxes, and patterns with horses, argonaut shells, spirals, and rosettes. The whole is colourful but more imaginative in idea than expert in execution. After about 1400, a series of small acropolis palaces was built, usually with a simple megaron hall, as at Tiryns, in Late Helladic III A. These palaces developed into almost grandiose complexes by the later 13th century, with lower courses of well-dressed limestone and painted floors, surrounded by workshops and storerooms. The descriptions of palaces in Homer are evidently based on memories of palaces such as these, and what Homer calls the megaron corresponds to the great hall. A small palace with mainland features was built at Phylakopi on Melos in the Cyclades, and a more regular form on massive foundations at Ayía Triadha near Phaistos in Crete. A shrine there whose floor was painted with fish and octopods looks forward to the painted floors of mainland palaces, with dolphins and octopods at Tiryns and octopods and fish at Pylos.

The palaces on the mainland had a system of keeping records that was similar to the Cretan one. The archive at Pylos, excavated by Blegen in 1939 and again after World War II, is the only extensive one found so far, but Thebes also produced tablets in some numbers, and there were smaller groups at Mycenae and Tiryns. These tablets reflect the same range of interest as those at Knossos: they consist of lists of palace personnel and of persons in outlying towns in professions such as bronzesmith, shepherd, cowherd, or tree cutter. There are lists of landowners, women and children, and priests and "slaves of the god," as well as records of agricultural income, of the preparation of perfumed oil, and of sacrifices of animals to the gods and offerings of oil and cloth at different parts of the Pylian province. Systems of landownership and tenancy were fairly complicated, and the palace kept a close eye on all dues and exchanges of goods. The archive at Thebes has records of trade with neighbours, in Euboea, or at places like Sicyon in the Peloponnese and of contacts with western Crete; the commercial interests of the Pylos district seem to have been more internal.

By the late 20th century, only three palace systems—at Tiryns, Mycenae, and Pylos—were well excavated and understood. The Theban palace may yet emerge; workshops, storerooms, and an arsenal have been found in probes under the modern town. Athens and Sparta may have had palaces, now lost; Dendra-Midea in the Argolid had impressive walls; Orchomenos in Boeotia had at least a small megaron with frescoes. Private houses are known both at the palace centres and in nonpalatial places, and some private houses, like those at Mycenae, maintained their own records in Linear B. There is a certain likeness all across Greece in architectural techniques, pottery, frescoes, ivory, and jewelry, but local autonomy and distinct variations in design and workshop styles also are evident. Gifts no doubt were exchanged among the principal centres, and there was at least a partial network of roadways connecting one centre to another; much trading must also have been coastal.

The Greek mainland in the 14th and 13th centuries was densely populated with towns and villages, and cemeteries confirm the numbers. The state was organized under a king, *wanax*, with a military leader, *rawaketa*, and troops with chariot officers attached for patrolling the borders; there also were naval detachments. The people had certain powers and a council. The towns were organized hierarchically under local officials, like the later "kings," *basileis*.

Hittite records

Eastward explorations. From the 15th century, the mainland Greeks explored eastward and replaced the Cretan settlers in such outpost towns as Tríanda on Rhodes or Miletus on the coast of Anatolia. There are Hittite records that apparently mention the maneuvers and political meddling of Greeks in coastal states; they refer to them under the name of Ahhíyawa, probably the equivalent to Homer's Achaeans at Troy. These records, from the 15th through the 13th century, are confirmed archaeologically by finds from the cemetery at Panaz Tepe near Phocaea in the north to Müskebi near Halicarnassus in the south. Panaz Tepe has warrior equipment, and apparently the soldiers took native wives, for the Greeks were buried while the Anatolians were cremated in the same small tholos tombs. Mycenaean pottery and imitations of it appeared at Troy itself from the 15th century onward. The renowned "Trojan War" may sum up a series of relationships and conflicts spanning the entire Bronze Age, since some of the archaic equipment described in the poems is actually found in 15th- and 14th-century Anatolia. There also was extensive trade with the Levantine coast and Cyprus, at least until all trade networks began to be disrupted after the Battle of Kadesh in the 13th century. The exports are far more visible than the reciprocal imports.

The end of the Bronze Age in the Aegean. From the middle of the 13th century, expensive fortification walls were constructed for the mainland palaces (except Pylos), which give testimony of tremendous skill in fitting large blocks of stone together without bonding, in designing sophisticated gates, and in protecting underground water supplies. At Tiryns the walls are marked by elegant setbacks, and at Mycenae the famous Lion Gate is ornamented with the sculpture of two lions, one on either side of a column. The gateway and walls on the Acropolis of Athens were also impressive, with postern gates and guard posts and roofed, sheltered water supplies, either from local springs or brought in by pipes. These walls may signal frictions between city-states such as marked classical Greece or represent a common fear of attack from enemies unknown to 20th-century investigators. It may be that the cost, in labour and hire, of these fortifications had serious effects on the economy. Yet the 13th-century palaces increased in size and complexity, their walls and floors being repainted. The tomb gifts did not decline in value, suggesting that local wealth was maintained, and, if the two columned tholos tombs at Mycenae, the Treasury of Atreus and the Tomb of Clytemnestra, were really built this late, as some scholars maintain, then dynastic resources were still potent. The palace workshops, controlling the production of blue glass paste jewelry, agate beads, or chariots and harness, also flourished until shortly before the end of the 13th century; these workshops had divine patrons, according to the texts. In the "private sector" outside the palace at Mycenae there was a shrine, apparently devoted to a popular cult that involved a fertility goddess, a sword goddess, and snakes.

Destruction of the mainland palaces

Shifts in populations. Toward the end of the 13th century, the mainland palaces were burned, possibly within a short time of each other; the exception was Thebes in the north, which may have received a destructive blow slightly earlier. Mycenae and Tiryns continued to be inhabited and indeed had very rich and energetic periods of pottery production and trade in the 12th century; Pylos was deserted, however, and Athens was inhabited but not wealthy. New centres, both of refuge and of independence, became conspicuous, such as Lefkandi on the inner shore of Euboea, south of Chalcis. The Cyclades, Crete, and, in the west, the Ionian islands such as Cephallenia experienced an increase in population. New expeditions eastward to Cyprus consisted of small groups who fortified military settlements around the coast. Anatolia may also have received new immigrants, as it had periodically since the 15th century, as far as Tarsus in Cilicia.

No completely satisfactory explanation for the collapse of the palace systems and the movements of populations has been found. Perhaps one must look toward a combination of factors such as climatic change and drought, harvest failure, starvation, epidemic, civic unrest, and resentment of palace taxes. Contributing factors may also have been the breaking off of trade with the east after the clash of the Hittites and Egyptians at the Battle of Kadesh earlier in the 13th century, the presence of roving piratical bands of both local peoples and immigrants around the coasts of the eastern Mediterranean (known in the Egyptian records as the Peoples of the Sea) who were hired as temporary allies by several states, and general frictions caused by

universally failing economies and alliances. At any rate, the stable states of the wealthy later Bronze Age, which had been bound by commercial exchanges and political alliances, gradually or swiftly collapsed into near chaos. By the end of the following Dark Age (lasting perhaps from 1100 to 1000 in some places, or 900 in others), new peoples had arrived and settled, as, for example, the Dorians in southern Greece and Crete and the southern Cyclades as far as Rhodes or the Phrygians in central Anatolia. Notable too is the fact that new late Hittite states had been formed in northern Syria at this time.

New foreign elements. Even before the end of the Bronze Age, there were occasional signs of new foreign elements in Greece, Crete, the islands, and Cyprus, such as exchanges of pottery and metalwork with Italy, Sardinia, the Balkans, and northern Greece and with regions like Epirus and other northern districts theretofore beyond the margins of the standard Mycenaean world. Dorian tribesmen as well as others may have moved into the weakened states and into the grazing lands to the south. They may have pushed the old inhabitants into flight or into isolated and linguistically separate hilltop areas like Arcadia. Occasionally single burials appear next to the family chamber tombs of traditional Mycenaean practice, along with an increase in cremation burial. Some alien gray pottery, pairs of long dress pins for securing women's untailored blanket garments, and the late introduction of iron and steel are further signs of new elements and evolving changes. Whether some of the "west Greek" elements in the new population were actually novel or had been present in Mycenaean society all along is obscured by the linguistic unity of the palace tablets; it may be that the Greek dialects that seem to have moved into place during the Dark Age, especially the Dorian dialects of the south, had been spoken but not written at an earlier stage by segments of the population.

The people of the Aegean Bronze Age. The Aegean populations after the Neolithic Period do not conform to a clear ethnic type. The men from small tribal organizations of early times seem to have chosen brides from outside the kin group, at distances from Anatolia to the Balkans and points south. Almost from the start one finds evidence of a variety of people—slender and stout, with round and long skulls, and of tall and medium height. Probably many of the ancient inhabitants of Greece and the islands looked as people in Greece do today—active, muscular, and of moderate height. From the evidence of the wall paintings, though these are often idealized, they seem largely to have had dark hair, dark or gray eyes, fine profiles, and slender figures. Detailed skeletal studies of burials in Grave Circle B at Mycenae have shown tall, rugged skeletons with large hands and feet, some arthritis and gallstones, and recurrent "family traits." The high average age at death—about 36 years—may reflect fighting careers, for which the men may have been socially selected, fed, and trained. Their generally superior physical condition in comparison to that of "commoners" was perhaps the result of a better diet from childhood onward.

Dress. Clay figurines of about 2000 from Crete show men wearing a narrow codpiece with a belt or loincloth and bare above the waist. This was to remain the basic fashion for Cretan men throughout the Bronze Age. Cretan women wore short-sleeved jackets that left the breasts bare and ankle-length flounced skirts, although shorter skirts to just below the knees are also attested. Marble figurines of men from the Cyclades assigned to the Early Bronze Age have belts and narrow codpieces like those of the Cretans. There is little evidence for dress on the mainland until the time of the Mycenae Shaft Graves in the 16th century. A considerable variety of dress is represented from that time onward throughout the Aegean area, but it is difficult to recognize fashions peculiar to the mainland. Tasseled shorts worn by men shown on the Mycenae lion-hunt dagger are also attested in Crete at the time. A group of Aegean envoys painted on the walls of the tomb of Rekhmire, vizier of the Egyptian pharaoh Thutmose III (ruled 1504–1450), are wearing large codpieces of a type fashionable in Crete in the Shaft Grave Period; thus they may have been Cretan envoys. A second group of envoys painted on the walls of the same tomb at a somewhat later date, however, are wearing kilts without codpieces, as worn by men in paintings at Knossos after the mainland conquest of Crete, about 1450. This might reflect mainland envoys going to Egypt after the conquest and wearing a different type of dress from the Cretans, but kilts of this kind appear to be represented in Crete both before and after the conquest.

In addition, curious scaly cloaks and long single-piece robes are among a variety of ritual garments in Crete. Linen was known in Crete by the beginning of the Bronze Age, and fragments of it were recovered from the Mycenae Shaft Graves; however, in Crete, at any rate, clothes were mostly, it seems, made of wool, and wall paintings show them woven with colourful and intricate designs, including pictures of animals and birds and even musical instruments. One of the dyes used was purple crushed from murex shells. Cretan men wore knee boots and sandals with upturned toes. Men with leggings or greaves are represented in wall paintings on the mainland. Caps of various kinds appear on the heads of men, and high, pointed hats and tiaras on those of women and of gods and goddesses or their priests. Clay figurines of the Early Palace Period show Cretan women with elaborate hair arrangements. Women put jewelry in their hair, including strings of beads. Necklaces, earrings, bracelets, and armlets were displayed by men as well as women. Sealstones were carried on strings around the neck or on the wrist. Cretan men normally left their hair long but were clean-shaven. Beards and mustaches are attested on the mainland in the Shaft Grave Period and later. Cretan dress

The frescoes at Thera show a wonderful variety of costumes, including the Minoan bodice-jacket, the flounced skirt or apron worn thigh-length or ankle-length, a one-piece tunic with rich borders, diaphanous veils, and a marvelous profusion of gold earrings, necklaces, collars, bracelets, and anklets, and rock crystal and carnelian beads. The men wear a kilt or a tunic or a loincloth; "peasants" may wear sheepskin cloaks; soldiers have long capes, tower shields, and boar's-tusk helmets.

Society. The early villages show few signs of economic disparity between families, although at times the presence of big houses in the later Neolithic Period indicates domination by chiefs. The island communities of the 3rd millennium are not yet well known, though signs of maritime trade are conspicuous and the grave gifts of marble idols point to organized religious rites and some wealth. The existence of fortified communities and two-story special houses on the mainland may indicate that communities contributed to their welfare and that they were ruled by a dynast. In Crete two types of early towns are known, a communal one, as at Myrtos, and one dominated by a big house or houses, as at Vasilikí. By the time of the Early Palaces, after 2000, it is clear that some governing power in several provinces was able to call upon extensive labour for the construction of buildings, granaries, and roads. The likenesses among the palaces, moreover, suggests that social systems across Crete were similar, perhaps dictated in form by certain religious behaviours. The palaces combined facilities for agricultural storage and for community displays and festivals, perhaps regulated by trained families and priestesses or priests. The palaces suggest a reciprocal relationship between the inhabitants and the surrounding villages. In mainland Greece, dynasties controlled fortified acropolis centres with outlying towns dependent on princes. This system is recorded extensively in Greek myths with Bronze Age origins, which tell of kings, princesses, and heroes from a few reigning families. During the last phase of Mycenaean culture and presumably during the Dark Age, the power of the old families was dispersed to lower local rulers, *basileis,* and the systems of councils of elders and village headman were maintained.

Trade. Foreign manufactures reaching the Aegean and especially Crete during the Bronze Age included Cypriot pottery, Mesopotamian and other Oriental cylinder seals, and Egyptian stone vases, ivories, and scarabs, while Cretan and eventually Mycenaean pottery is found in Egypt and elsewhere in the Levant.

By the 14th and 13th centuries, Mycenaean pottery is found densely in the Levant; it is often accompanied by Cypriot pottery as though carried in Cypriot or Syrian ships. Mycenaean pottery not mixed with Cypriot pottery is found in Anatolia from Troy to Tarsus. Because there is almost nothing on the mainland in return, one may suppose that trade was carried on in archaeological invisibles, such as food, textiles, copper ores, and perhaps slaves or war captives (some are attested in the Linear B texts). Mycenaeans may also have exported technology, such as weapon making, or mercenaries. Crete and the mainland had to import tin for bronze, probably from Anatolia, and both used copper ores from Cyprus and other sources. Minoan contact seems to have reached Sicily and Sardinia, and metal ingots may have been brought back from the west. Silver-lead was produced in the Cyclades and Attica. The Kaş Ulu Burun shipwreck shows an extensive trade in glass ingots, often cobalt blue, as well. Ostrich eggs and stone for making vases were among imports to Crete from Egypt, and ivory came from there or from Syria. Amber from the Baltic reached the mainland in some quantity during the Shaft Grave Period and later but is rarely found in Crete. Exports from the Aegean may have included woolen goods, olive oil, and timber, as well as silver. In Crete, at any rate, foreign trade may have been largely under palace control, but a class of private merchants engaged in overseas commerce no doubt existed in the Aegean.

Transport. Ships with a mast and square sail in addition to oars or paddles were used in the Aegean from the Early Bronze Age. On land, goods were no doubt transported by pack animals or on poles slung between bearers; this principle was also adopted for passenger chairs, of which there are clay models. A model of a four-wheeled cart from Crete is datable to about 2000 or earlier. The wheels of such carts were evidently solid, and the carts were no doubt drawn by oxen. Horses may have been ridden in Crete by then, as they seem to be depicted on early Cretan seals. These horses could have come from the east, but a different breed was introduced into the mainland from the north at the beginning of the Middle Bronze Age, about 2000. The light spoke-wheeled chariot drawn by horses appears to have developed in Syria or northern Mesopotamia early in the 2nd millennium, but it spread rapidly throughout the Middle East because of its usefulness in war. Chariots are depicted on tombstones of the Mycenae Shaft Graves and on Cretan seals before the time of the mainland conquest, about 1450. Apart from warfare, they were used in the Aegean for hunting and probably for travel. During the latter part of the Bronze Age, terraces were built to support roads wide enough for wheeled vehicles both in Crete and on the mainland. Such roads were carried across streams on bridges, examples of which have survived in the region of Mycenae.

Carts and chariots

Warfare. Short daggers of types derived from Syria were in use in the Aegean during the Bronze Age. Long rapiers, evolved from those in Crete, are found on the mainland by the time of the Mycenae Shaft Graves in the 16th century BC.

The traditional armour of the Shaft Grave Period—a shield shaped in the figure eight or a tower shield, a helmet often reinforced with boars' tusks, a thrusting spear, and a sword on a baldric in a tasseled scabbard—appears also in the Thera naval fresco and in the epics behind Homer's *Iliad.* Charioteers apparently wore a bronze tunic of thonged plates, sketched on the Knossos tablets and found in a chamber tomb at Dendra in the Argolid. Linen greaves appear in frescoes, and bronze greaves in graves. There were bronze wrist guards for archers. Many soldiers may have preferred quilted, padded protection in the summer because of the heat.

Short swords adapted for cutting as well as thrusting began to appear in the following century and may have been developed in connection with chariot warfare. Bronze armour and small, round shields more serviceable in chariots replaced the old Cretan body shields at approximately the same time. Bows and slings were probably used everywhere in the Aegean area, but, whereas arrowheads of flint and obsidian are found on the mainland, they are virtually unknown in Crete, where arrows may have been tipped with bone or wood until the appearance of bronze arrowheads in the 15th century. Settlements on the mainland and in the Cyclades were defended by walls from the Early Bronze Age onward, and the town at Mallia in Crete appears to have been protected by a wall during the period of the Early Palaces; but, by the time of the Late Palaces, Cretan towns may have been unwalled. Faience inlays of the 17th century from Knossos, however, seem to show an attack on a walled town such as that depicted on a silver-relief vase from the Mycenae Shaft Graves. The attraction of the theme of the city by the sea, with vignettes of war and peace, landscape and water, is also apparent in the Thera naval fresco and the Master Sealing of Chania in western Crete, which shows a youth lording it over the rooftops of a town. Methods of warfare had become highly developed by the end of the Bronze Age, with improved weapons, complex and well-designed fortifications, extensive use of chariots, and warships with rams.

Development of fortified towns



Warrior armed with a figure-of-eight shield and boar's-tusk helmet, ivory relief, *c.* 1400–1200 BC, from Delos. In the Archaeological Museum, Delos, Greece.
Cliche, Ecole Francaise d'Archeologie, Athenes

Religion. Little is known about religion in the Cyclades and on the mainland before the period when they came under strong Cretan influence. An open-air sanctuary filled with marble figurines on the island of Kéros (Káros) is assignable to the Early Bronze Age. In Crete during the Early Palace Period, there were many open-air sanctuaries on the tops of hills and mountains. Some of these had small shrines in them, and shrines with one or more rooms and benches for offerings and cult statues are found in the countryside and in the towns in Crete. Parts of the palaces and of large houses there were also set apart for cult. Shrines not unlike Cretan ones existed in settlements in the Cyclades and on the mainland in the Late Bronze Age; however, hilltop sanctuaries are not well attested there, and most of those in Crete appear to have gone out of use after the mainland conquest, about 1450. Caves also were used as sanctuaries in Crete, and cults in some of these persisted until the end of the Bronze Age and later.

The chief deity everywhere in the Aegean during the Bronze Age was evidently a goddess. Perhaps there were several goddesses with different names and attributes. The extant texts refer to a Potnia ("Lady" or "Mistress"), to whom they give several epithets like "horse" or "grain." Most mainland palaces have paintings of processions in

which people bring gifts to a goddess. On Thera, frescoes show girls picking saffron crocus and offering it in baskets to a seated goddess. Clay statues of goddesses, often with upraised arms and attributes such as horns of consecration, doves, snakes, or poppies have been found in Crete; these range in date from the 14th to the early 12th century, providing evidence of a strong tradition. A shrine with large clay goddesses, which once were stuccoed and painted, existed at Ayía Iríni on the island of Ceos, and a smaller, later one at Phylakopi on Melos, with both male and female figurines. The shrine at Mycenae seems to have been devoted to powers of grain and the sword. A later shrine at Tiryns had small clay goddesses with upraised arms. Many cult statues may have been made of wood, and mythic traditions of simple wooden logs or planks (xoana) dropping from heaven or being found in thickets have become attached to several later sanctuaries.

The texts show a more elaborate set of divinities than do the surviving idols, with many later Greek divinities already in place, including Zeus, Poseidon, Athena, Artemis, Ares, Hermes, and Dionysus. The Cretan birth goddess Eleuthia and war goddess Eyno were transmitted to the mainland Greeks, and natural forces, like the winds, were occasionally worshiped. There can be no doubt about the continuity of religions and cult from the Late Bronze Age into later Greek times, as well as of the language itself. Some divinities, like the female Zeus and the female Poseidon figures known at Pylos, do not reappear in later times, however. The culture was reshaping itself as it passed from generation to generation.

The normal gifts to divinities were scented oils, textiles, and, in Greece at least, animal sacrifice of cattle, sheep, and pigs. The burial of a horse or a dog may either signify a sacrifice or simply express the attachment between the animal and its master. Two ideas about the realm of death existed, a rarer one of an overseas Elysian paradise where the dead were restored to a new life of bodily blessed ease and a more common one, transmitted in the epic tradition, of a dark underground realm (Hades) inhabited by weak shades with poor memories. These two ideas, representing the Cretan and the Mycenaean tradition, were not fused but survived in separate sets of songs and tales.

(M.S.F.H./E.D.T.V.)

Ancient Greek civilization

THE EARLY ARCHAIC PERIOD

The post-Mycenaean period and Lefkandi. The period between the catastrophic end of the Mycenaean civilization and about 900 BC is often called a Dark Age. It was a time about which Greeks of the Classical age had confused and actually false notions. Thucydides, the great ancient historian of the 5th century BC, wrote a sketch of Greek history from the Trojan War to his own day, in which he notoriously fails, in the appropriate chapter, to signal any kind of dramatic rupture. (He does, however, speak of Greece "settling down gradually" and colonizing Italy, Sicily, and what is now western Turkey. This surely implies that Greece was settling down after something.) Thucydides does indeed display sound knowledge of the series of migrations by which Greece was resettled in the post-Mycenaean period. The most famous of these was the "Dorian invasion," which the Greeks called, or connected with, the legendary "return of the descendants of Heracles." Although much about that invasion is problematic—it left little or no archaeological trace at the point in time where tradition puts it—the problems are of no concern here. Important for the understanding of the Archaic and Classical periods, however, is the powerful belief in Dorianism as a linguistic and religious concept. Thucydides casually but significantly mentions soldiers speaking the "Dorian dialect" in a narrative about ordinary military matters in the year 426: this is a surprisingly abstract way of looking at the subdivisions of the Greeks, because it would have been more natural for a 5th-century Greek to identify soldiers by cities. Equally important to the understanding of this period is the hostility to Dorians, usually on the part of Ionians, another linguistic and religious subgroup, whose most famous city was Athens. So extreme was this hostility that Dorians were prohibited from entering Ionian sanctuaries; extant today is a 5th-century example of such a prohibition, an inscription from the island of Paros.

The importance of the Dorian invasion

Phenomena such as the tension between Dorians and Ionians that have their origins in the Dark Age are a reminder that Greek civilization did not emerge either unannounced or uncontaminated by what had gone before. The Dark Age itself is beyond the scope of this article. One is bound to notice, however, that archaeological finds tend to call into question the whole concept of a Dark Age by showing that certain features of Greek civilization once thought not to antedate about 800 BC can actually be pushed back by as much as two centuries. One example, chosen for its relevance to the emergence of the Greek city-state, or polis (see below), will suffice. In 1981 archaeology pulled back the curtain on the "darkest" phase of all, the Protogeometric Period (*c.* 1075–900 BC), which takes its name from the geometric shapes painted on pottery. A grave, rich by the standards of any period, was uncovered at a site called Lefkandi on Euboea, the island along the eastern flank of Attica (the territory controlled by Athens). The grave, which dates to about 1000 BC, contains the (probably cremated) remains of a man and a woman. The large bronze vessel in which the man's ashes were deposited came from Cyprus, and the gold items buried with the woman are splendid and sophisticated in their workmanship. Remains of horses were found as well; the animals had been buried with their snaffle bits. The grave was within a large collapsed house, whose form anticipates that of the Greek temples two centuries later. Previously it had been thought that these temples were one of the first manifestations of the "monumentalizing" associated with the beginnings of the city-state. Thus this find, and those made in a set of nearby cemeteries in the years before 1980 attesting further contacts between Egypt and Cyprus between 1000 and 800 BC, are important evidence. They show that one corner of one island of Greece, at least, was neither impoverished nor isolated in a period usually thought to have been both. The difficulty is to know just how exceptional Lefkandi was, but in any view it has revised former ideas about what was and what was not possible at the beginning of the 1st millennium BC.

Colonization and city-state formation. The first "date" in Greek history is 776 BC, the year of the first Olympic Games. It was computed by a 5th-century-BC researcher called Hippias. This man originally came from Elis, a place in the western Peloponnese in whose territory Olympia itself is situated. This date and the list of early victors, transmitted by another literary tradition, are likely to be reliable, if only because the list is so unassuming in its early

The first Olympic Games

© British School at Athens



Centaur, terra-cotta, 10th century, from the site of Lefkandi, Euboea, Greece.



Ancient Greece.
From W. Shepherd, *Historical Atlas*; Barnes and Noble, Inc.

reaches. That is to say, local victors predominate, including some Messenians. Messene lost its independence to neighbouring Sparta during the course of the 8th century, and this fact is an additional guarantee of the reliability of the early Olympic victor list: Messenian victors would hardly have been invented at a time when Messene as a political entity had ceased to exist. Clearly, then, record keeping and organized activity involving more than one community and centring on a sanctuary, such as Olympia, go back to the early 8th century. (Such competitive activity is an example of what has been called "peer-polity interaction.") Records imply a degree of literacy, and here too the tradition about the 8th century has been confirmed by recent finds. A cup, bearing the inscription in Greek in the Euboean script "I am the cup of Nestor," can be securely dated to before 700 BC. It was found at an island site called Pithekoussai (Ischia) on the Bay of Naples.

The Euboeans, whose early overseas activity has already been remarked upon in connection with the discoveries at Lefkandi, were the prime movers in the organized and recorded phase of Greek colonization. (This point can be taken as absolutely certain because archaeology supports the literary tradition of the Roman historian Livy and others: Euboean pottery has been found both at Pithekoussai to the west and at the Turkish site of Al-Mina to the east.) This organized phase began in Italy *c.* 750 and in Sicily in 734 BC; its episodes were remembered, perhaps in writing, by the colonies themselves. The word "organized" needs to be stressed, because various considerations make it necessary to push back beyond this date the beginning of Greek colonization. First, it is clear from archaeological finds, such as the Lefkandi material, and from other new evidence that the Greeks had already, before 750 or 734, confronted and exchanged goods with the inhabitants of Italy and Sicily. Second, Thucydides says that Dark Age Athens sent colonies to Ionia, and archaeology bears this out—however much one discounts for propagandist exaggeration by the imperial Athens of Thucydides' own time of its prehistoric colonizing role. However, after the founding of Cumae (a mainland Italian offshoot of the island settlement of Pithekoussai) *c.* 750 BC and of Sicilian Naxos and Syracuse in 734 and 733, respectively, there was an explosion of colonies to all points of the compass. The only exceptions were those areas, such as pharaonic Egypt or inner Anatolia, where the inhabitants were too militarily and politically advanced to be easily overrun.

One may ask why the Greeks suddenly began to launch these overseas projects. It seems that commercial interests, greed, and sheer curiosity were the motivating forces. An older view, according to which Archaic Greece exported its surplus population because of an uncontrollable rise in population, must be regarded as largely discredited. In the first place, the earliest well-documented colonial operations were small-scale affairs, too small to make much difference to the situation of the sending community (the "metropolis," or mother city). That is certainly true of the colonization of Cyrene, in North Africa, from the island of Thera (Santorin): on this point an inscription has confirmed the classic account by the 5th-century Greek historian Herodotus. In the second place, population was not uncontrollable in principle: artificial means such as infanticide were available, not to mention more modern

Reasons for the colonization movement

techniques like contraception. Considerations of this kind much reduce the evidential value of discoveries establishing, for example, that the number of graves in Attica and the Argolid (the area round Argos) increased dramatically in the later Dark Age or that there was a serious drought in 8th-century Attica (this is the admitted implication of a number of dried-up wells in the Athenian agora, or civic centre). In fact, no single explanation for the colonizing activity is plausible. Political difficulties at home might sometimes be a factor, as, for instance, at Sparta, which in the 8th century sent out a colony to Taras (Tarentum) in Italy as a way of getting rid of an unwanted half-caste group. Nor can one rule out simple craving for excitement and a desire to see the world. The lyric poetry of the energetic and high-strung poet Archilochus, a 7th-century Parian involved in the colonization of Thasos, shows the kind of lively minded individual who might be involved in the colonizing movement.

So far, the vague term "community" has been used for places that sent out colonies. Such vagueness is historically appropriate, because these places themselves were scarcely constituted as united entities, such as a city, or polis. For example, it is a curious fact that Corinth, which in 733 colonized Syracuse in Sicily, was itself scarcely a properly constituted polis in 733. (The formation of Corinth as a united entity is to be put in the second half of the 8th century, with precisely the colonization of Syracuse as its first collective act.)

The name given to polis formation by the Greeks themselves was *synoikismos,* literally a "gathering together." *Synoikismos* could take one or both of two forms—it could be a physical concentration of the population in a single city or an act of purely political unification that allowed the population to continue living in a dispersed way. The classic discussion is by Thucydides, who distinguishes between the two kinds of *synoikismos* more carefully than do some of his modern critics. He makes the correct point that Attica was politically synoecized at an early date but not physically synoecized until 431 BC when Pericles as part of his war policy brought the large rural population behind the city walls of Athens. A more extreme instance of a polis that was never fully synoecized in the physical sense was Sparta, which, as Thucydides elsewhere says, remained "settled by villages in the old Greek way." It was an act of conscious arrogance, a way of claiming to be invulnerable from attack and not to need the walls that Thucydides again and again treats as the sign and guarantee of civilized polis life. The urban history of Sparta makes an interesting case history showing that Mycenaean Sparta was not so physically or psychologically secure as its Greek and Roman successors. The administrative centre of Mycenaean Sparta was probably in the Párnon Mountains at the excavated site of the Menelaion. Then Archaic and Classical Sparta moved down to the plain. Byzantine Sparta, more insecure, moved out of the plain again to perch on the site of Mistra on the opposite western mountain, Taygetos. Finally, modern Sparta is situated, once again peacefully and confidently, on its old site on the plain of the river Eurotas.

The beginnings of the polis

The enabling factors behind the beginnings of the Greek polis have been the subject of intense discussion. One approach connects the beginnings of the polis with the first monumental buildings, usually temples like the great early 8th-century temple of Hera on the island of Samos. The concentration of resources and effort required for such constructions presupposes the formation of self-conscious polis units and may actually have accelerated it. As stated above, however, the evidence from Lefkandi makes it hard to see the construction of such monumental buildings as a sufficient cause for the emergence of the polis, a process or event nobody has yet tried to date as early as 1000 BC. Another related theory argues that the birth of the Greek city was signaled by the placing of rural sanctuaries at the margins of the territory that a community sought to define as its own. This fits admirably a number of Peloponnesian sanctuaries; for instance, the temple complex of Hera staked out a claim, on the part of relatively distant Argos, to the plain stretching between city and sanctuary, and the Corinthian sanctuary on the promontory of Perachora, also dedicated to Hera, performed the same function. Yet there are difficulties. It seems that the Isthmia sanctuary, which at first sight seems a good candidate for another Corinthian rural sanctuary, was already operational as early as 900 BC, in the Protogeometric Period, and this date is surely too early for polis formation. Nor does the theory easily account for the rural temple of the goddess Aphaea in the middle of Aegina. The sanctuary is admittedly a long way from the town of Aegina, but Aegina is an island, and there is no obvious neighbour against whom territorial claims could plausibly have been asserted. Finally, a theory that has to treat Athens and Attica as in every respect exceptional is not satisfactory: the relation of Athens to Eleusis (the sanctuary on its western margins and the only possible candidate for a "rural sanctuary" in the required sense) is very different from that of the Peloponnesian cases already considered above. Eleusis was far more than a mere "outpost"; there was a tradition that it had once been an independent state.

A third theory attacks the problem of the beginnings of the polis through burial practice. In the 8th century (it is said) formal burial became more generally available, and this "democratization" of burial is evidence for a fundamentally new attitude toward society. The theory seeks to associate the new attitude with the growth of the polis. There is, however, insufficient archaeological and historical evidence for this view (which involves an implausible hypothesis that the process postulated was discontinuous and actually reversed for a brief period at a date later than the 8th century). Moreover, it is vulnerable to the converse objection as that raised against the second theory: the evidence for the third theory is almost exclusively Attic, and so, even if it were true, it would explain Athens and only Athens.

Fourth, one may consider a theory whose unspoken premise is a kind of "geographic determinism." Perhaps the Greek landscape itself, with its small alluvial plains often surrounded by defensible mountain systems, somehow prompted the formation of small and acrimonious poleis, endlessly going to war over boundaries. This view has its attractions, but the obvious objection is that, when Greeks went to more open areas such as Italy, Sicily, and North Africa, they seem to have taken their animosities with them. This in turn invites speculations of a psychologically determinist sort; one has to ask, without hope of an answer, whether the Greeks were naturally particularist.

A fifth enabling factor that should be borne in mind is the influence of the colonizing movement itself. The political structure of the metropolis, or sending city, may sometimes have been inchoate. The new colony, however, threatened by hostile native neighbours, rapidly had to "get its act together" if it was to be a viable cell of Hellenism on foreign soil. This effort in turn affected the situation in the metropolis, because Greek colonies often kept close religious and social links with it. A 4th-century inscription, for instance, attests close ties between Miletus and its daughter city Olbia in the Black Sea region. Here, however, as so often in Greek history, generalization is dangerous; some mother-daughter relationships, like that between Corinth and Corcyra (Corfu), were bad virtually from the start.

A related factor is Phoenician influence (related, because the early Phoenicians were great colonizers, who must often have met trading Greeks). The Phoenician coast was settled by communities similar in many respects to the early Greek poleis. It is arguable that Phoenician influence, and Semitic influence generally, on early Greece has been seriously underrated.

Theories such as these are stimulating and may each contain a particle of truth. The better position, however, is that generalization itself is as yet premature; in particular, archaeologically based theoretical reconstructions need much more refining. All one can say in summary is that in roughly the same period—namely, the 8th century—a number of areas, such as Corinth and Megara, began to define their borders, deny autonomy to their constituent villages, and generally act as separate states. Attica's political synoecism, which occurred a little earlier, was complete perhaps around 900. Tempting though it is

to seek a single explanatory model for these very roughly contemporaneous processes, one should perhaps allow that different paths of development were followed in different areas, even in areas next door to each other. After all, the Archaic and Classical histories of mighty democratic imperial Athens, of the miserable polis of Megara which nevertheless colonized Byzantium, of wealthy, oligarchic Corinth, and of federal Boeotia were all very different even though Athens, Megara, Corinth, and Boeotia were close neighbours.

Religious leagues (amphictyonies)

One is perhaps on firmer ground when one examines the evidence for prepolis aggregations of larger units, often religious in character. There are a number of such associations whose origins lie in the Dark Age and whose existence surely promoted some feeling of local and particularist identity among the participants. The Ionians in Anatolia formed themselves into a confederation of 12 communities, the Ionian Dodecapolis, with a common meeting place; and there were comparable groupings among the Dorian Greeks of Anatolia and even among the Carians (partially Hellenized non-Greeks) in the same part of the world. The central location for such organizations was characteristically small and insignificant. One poorly attested but intriguing early Archaic league was the "Calaurian Amphictyony" (an amphictyony was a religious league of "dwellers round about"). Calauria, the small island now called Póros, was not a place of any consequence in itself, but the league's seven members included Athens and Aegina, two major Greek poleis. The most famous and enduring such amphictyony, however, was the one that, originally from a distance, administered the affairs of the sanctuary of Delphi in central Greece. This sanctuary contained the most famous, though not the oldest, Greek oracle (the oldest was at Dodona); oracles were a mechanism by which divinely inspired utterances were produced in answer to specific questions. Finally, it is worth noting an adventurous suggestion that Lefkandi itself might have been the centre of some kind of religious amphictyony, but, if so, this would be an exception to the principle that religious centres tended themselves to be insignificant, however mighty their participating members.

Early Archaic Greek civilization. *The sources.* Before attempting to characterize Archaic Greece, one must admit candidly that the evidence is unsatisfactory. Only for Athens is anything like a proper political tradition known, and Athens' development toward the democracy of the 5th century was amazingly and untypically rapid by comparison with other states, many of which never became democratic at all. A tiny but salutary scrap of evidence makes this point: Thucydides in Book 2 of his *History of the Peloponnesian War* casually mentions a man called Evarchus as "tyrant" of a small northwestern Greek polis called Astacus in the 420s BC. But for this chance mention, one would never have guessed that tyranny could have existed or persisted in such a place so late or so long. Another difficulty is that, while a fair amount about the social structure of Classical Athens is known, some of it must go back to Archaic times; just how much is disputed.

The nature of the evidence

There is a further complication. In both the political and the social spheres, one has to reckon—chiefly at Athens, but elsewhere too—with "invented tradition," a distorting element for which proper allowance is only now beginning to be made. Thus it seems that not just Lycurgus, the famous Spartan lawgiver (whose historicity was doubted even in antiquity), but even a reforming figure like Solon of Athens, who certainly existed in the 6th century and large fragments of whose poetry still survive, was in some respects what anthropologists call a "culture hero." Much was projected onto him anachronistically or just wrongly, and reformers in later generations established their credentials by claiming (if they were reactionaries) that they were trying to "get back to Solon" or (if they were democrats) that Solon was their founding father. Such errors should not induce too much pessimism: at Athens at least, individual aristocratic families preserved oral traditions, which affected the later literary records in ways that can be properly understood with the help of anthropological analogy. That is to say, not all the evidence so preserved is unusable, but it needs handling in special ways.

It has even been argued that social life, too, was creatively manipulated. Later Greek cities contained, alongside such transparent political institutions as the Popular Assembly and the Deliberative ("Probouleutic") Council, a more opaque set of institutions, ostensibly based on kinship groupings. The biggest and most basic of these groupings were the phylae, or "tribes," according to which the citizen body was subdivided. Thus all Dorian states had the same three tribes, and there were four Ionian tribes (although Ionian states were less conservative than Dorian, and one finds among them a greater readiness to innovate; late 6th-century Athens, for example, switched from a four-tribe hereditary system of citizenship to a 10-tribe one based on simple residence as well as descent). Smaller subdivisions were the phratry, a word connected with a philological root meaning "brother," and the *genos,* a smaller cluster of families (*oikoi*). The existence of these groupings in historical times is beyond question; one finds them controlling citizen intake (as in the so-called "Demotionidai" inscription from the Attic village of Decelea, datable to as late as the early 4th century BC) and entering into complicated property arrangements. What has become a matter of debate, however, is the question of just how old they actually were. According to the most skeptical view, the whole apparatus of tribe and *genos* was an invention without any Dark Age history to legitimate it. This view, which rests partly on the near absence of the relevant kinship terminology in Homer, is not ultimately convincing in its hypothesis of a kind of complicated collective fraud on posterity. Yet it is right to allow for an element of conscious antiquarianism at certain periods (the 320s in Athens being one), which may well have affected specific traditions.

Society and values. The world of the colonizing states was aristocratic in the sense that a small number of exclusive clans within cities monopolized citizenship and political control. At Corinth, for example, citizenship was confined to the adult males of a single clan, the Bacchiadae. They perhaps numbered no more than a couple of hundred. At Athens there was a general class of Eupatridae, a word that just means "People of Good Descent"—*i.e.,* aristocrats. (The word may have had a simultaneous but narrower application to one single *genos.* This, however, is disputed, and, in any event, that hypothetical family was only one among many privileged *genē.* The case, therefore, is not analogous to that of the Bacchiadae.) It is unlikely that the Eupatridae were as rigidly defined as the Bacchiadae, and the negative tradition that Solon in the early 6th century deprived them of their exclusive claim to political office may just be the excessively formal and precise way in which later ancient commentators described a positive change by which power was made more generally available than it had been before (see below).

Bacchiadae and Eupatridae

With regard to these same early Archaic times one hears, for example, in the poetry of the 7th-century Boeotian Hesiod, of control, sometimes oppressively exercised, by *basileis* (singular *basileus*). This word is usually translated as "kings," and such titles as the Athenian *basileus* (an official, or archon, with a defined religious competence, conveniently but less correctly called the *archon basileus* by modern scholars) are then explained as survivals of an age of monarchy. This account in terms of fossilization certainly eases the awkwardness of explaining why, for instance, the wife of the *archon basileus* was held to be ritually married to the god Dionysus. The very existence of kingship in Geometric (as opposed to Mycenaean) Greece, however, has been challenged, and a case has been made (though not universally accepted) for seeing most of these Archaic *basileis* not as kings in any sense but as hereditary nobles. In the latter case, there is no great difference between these *basileis* and such aristocrats as the Bacchiadae.

Life inside the Archaic Greek societies ruled by such families can be reconstructed only impressionistically and only at the top of the social scale; the evidence, to an extent unusual even in Greco-Roman antiquity, is essentially elitist in its bias. Aristocratic values were transmitted both vertically, by family oral traditions, and horizontally, by means of a crucial institution known as the symposium,

or feast, for which (many literary scholars now believe) much surviving Archaic poetry was originally written. Perhaps much fine painted pottery was also intended for this market, though the social and artistic significance of such pottery is debated. Some scholars insist that the really wealthy would at all times have used gold and silver vessels, which, however, have not survived in any numbers because they have long ago been melted down. Symposia were eating and dining occasions with a strong ritual element; their existence is reflected in the marked emphasis, in the Homeric poems, on ostentatious feasting and formal banqueting as assertions of status (what have been called "feasts of merit"). Thus Sarpedon in Homer's *Iliad* reminds Glaucus that both of them are honoured with seats of honour and full cups in Lycia and with land (a sacred precinct, or temenos) to finance all the feasting. Symposia were confined to males (a reminder of the military ethos so prevalent in Homer); although when the institution was introduced, along with the vine, to Etruria—where much of the visual evidence comes from—it changed its character and became open to both sexes. The Greek symposium proper can be seen as an instrument of social control; it is a more tangible unit of social organization, and one with better-attested Homeric antecedents, than the problematic *genē* or phratries discussed above. There is much to be said for this suggestion, but the evidence needs to be systematically collated and interpreted.

Symposia and gymnasia

In Classical times, strong homosexual attachments were another way in which values were inculcated, passed on by the older man (the *erastes*) to the younger *eromenos,* or beloved. The gymnasium was the venue where such relationships typically developed. As with the symposium, there was an almost ritual element to it all; certain gifts—such as, for example, the gift of a hare—were thought especially appropriate. The date, however, at which Greek homosexuality became a central cultural institution is problematic; it is notoriously absent from the Homeric poems, a fact that some scholars explain as being the result of poetic reticence. The more plausible view is that homosexuality was in some way connected with the rise of the polis and was part of what has been called the "8th-century renaissance." If so, Homer's silence is after all significant: he does not mention it because in his time it was not yet important.

Both symposia and gymnasia in different ways mirrored or were preparatory to warfare (see below). Interpolis athletic competitions (such as the Olympic Games) are another reflection of warfare. Epinician poetry of the Classical period (that is, "victory poetry" like that of Pindar, whose epinician odes celebrate the athletic victories of aristocratic individuals) constantly uses the language of war, fighting, and victory. Indeed, one influential view of organized athletic competitions is that they are a restructuring of the instinct to hunt and kill.

With the great athletic festivals, which brought Greeks together at set intervals of years to Olympia and later to Delphi, Nemea, and the Isthmus (the four great Panhellenic, or "all-Greek," games), one passes from the internal organization of individual Greek societies to their interrelationships. Two kinds of powerful interrelationship have already been noted—that between colonizing city and daughter city and the shared membership of an amphictyony.

More generally, the basic institution in intercity relationships was that of "guest-friendship," or *xenia.* This was another area where ritual elements were present to such a marked degree that the whole institution has been called "ritualized friendship." The same aristocrats who drank and heard poetry together inside their own communities naturally expected to find comparable groups inside other states. They cemented their ties, which had perhaps been formed on initially casual or trading visits, with formal relationships of *xenia.* At some point quite early in the Archaic period this institution developed into something still more definite, the *proxenia. Proxenoi* were citizens of state A living in state A who looked after the interests of citizens of state B. The status of *proxenos* was surely in origin hereditary, but by Thucydides' time one hears of "voluntary *proxenoi*" (*etheloproxenoi*). The antiquity of the basic

Xenia and *proxenia*

institution is not in doubt, however much the 5th-century Athenian empire may have exploited and reshaped it for its own political convenience; a 7th-century inscription from the island of Corcyra mentioning a *proxenos* from Locris is the earliest attestation of the institution.

Another way of institutionalizing relationships between the nationals of different states was *epigamia,* an arrangement by which the offspring of marriage were treated as citizens of the wife's polis if the husband settled there; and so was the husband. Athens, for example, granted *epigamia* to Euboea as late as the 5th century, a time when Athenian citizenship was fiercely protected. There are still earlier instances: usually one hears of *epigamia* when for one reason or another it was being suspended or denied. Thus, there was an early arrangement between the islands of Andros and Paros, which, Plutarch says, ended when relations went sour. More interesting is the statement, again by Plutarch, that there was no intermarriage between members of two of the villages, or demes, of Attica, Pallene and Hagnous. Far from being evidence that these places were somehow originally separate states, the prohibition was more like a ban on endogamy: in other words, the two communities were regarded—like members of a family—as being too close to be allowed to intermarry. Thus, both marriage itself and prohibition of marriage were ways of defining the relations between communities, including communities within a single large state like Attica, and of keeping those relations friendly.

The chief vehicle of interaction among poleis, however, was through warfare and through the formal suspension or renunciation of warfare by means of heavily ritualized treaties (one of the most common words for such a treaty is *spondai,* which literally means "libations" to the guaranteeing gods). The earliest surviving inscriptional peace treaty "for all time" dates from the 6th century and was found at Olympia. Nonetheless, there were surely agreements to limit warfare over strips of boundary land before that date. Archaeology may offer unexpected help in this matter: it is possible and plausible that some frontier zones were by tacit or explicit agreement left fallow. One such zone seems to have been the remote Skourta plain, which separates part of northern Attica from Boeotia; preliminary surface survey (*i.e.,* the estimation of settlement patterns by gathering of potsherds) carried out in and after 1985 suggests that it was left—perhaps deliberately—uncultivated in the Archaic period.

An important landmark in interstate military relations of the kind considered here was the Lelantine War. It was the earliest Greek war (after the mythical "Trojan War") that had any claim to be considered "general," in the sense that it involved distant allies on each side. Fought in perhaps the later 8th century between the two main communities of Euboea, Chalcis and Eretria, it took its name from the fertile Lelantine Plain, which separates the cities and includes the site of Lefkandi. (It is an interesting modern suggestion that Lefkandi itself is the site of Old Eretria, abandoned about 700 BC in favour of the classical site Eretria at the east end of the plain, perhaps as a consequence of Eretria's defeat in the war. This theory, however, needs to account for Herodotus' statement that at the early 6th-century entertainment of the suitors of Cleisthenes of Sicyon [see below] there was one Lysanias from Eretria, "then at the height of its prosperity.") Other faraway Greek states were somehow involved in the war; on this point Thucydides agrees with his great predecessor Herodotus. Thus Samos supported Chalcis and Miletus, Eretria. Given Euboean priority in colonization, it is natural to suppose that the links implied by the traditions about the Lelantine War were the result of Euboean overseas energy, but that energy would hardly have turned casual contacts into actual alliances without a preliminary network of guest-friendships. Whether the oracle at Delphi took sides in the war, as a modern speculation has it, is less certain, though there is no doubt that by some means wholly mysterious to the 20th century Delphi often provided updated information about possible colonial sites and even (as over Cyrene) gave the original stimulus to the colonizing act.

The Lelantine War

One can be more confident in denying the thoroughly

anachronistic notion that the Lelantine War shows the existence of "trade leagues" at this early date. Religious amphictyonies are one thing and trade leagues another; the evidence, such as it is, suggests that early trade was carried on by entrepreneurial aristocratic individuals, who no doubt exploited their guest-friendships and formed more such friendships during their travels. It is true, however, that such individuals tended to come from areas where arable land was restricted, and to this extent it is legitimate to speak in a generic way of those areas as having in a sense a more commercially minded population than others. One example of such an area is the Lelantine Plain, an exceptionally good piece of land on a notably barren and mountainous, though large, island. Herodotus described one such trader from the later Archaic period, Sostratus of Aegina, a man of fabulous wealth. Then in the early 1970s a remarkable inscription was found in Etruria—a dedication to Apollo in the name of Sostratus of Aegina. This discovery revealed that the source of his wealth was trade with Etruria and other parts of Italy. Aegina is an island whose estimated Classical population of about 40,000 was supported by land capable of supporting only about 4,000. One may quarrel with the first figure as too large and the second figure as too pessimistic (it makes insufficient allowance for the possibilities of highly intensive land use). Even after adjustment, however, it is clear that Aegina needed to trade in order to live. It is not surprising to find Sostratus' home city of Aegina included among the Greek communities allowed to trade at Naukratis in pharaonic Egypt; this arrangement is described by Herodotus, and the site has been explored archaeologically. Aegina was the only participating city of Greece proper, as opposed to places in the eastern Aegean.

THE LATER ARCHAIC PERIODS

The rise of the tyrants. Dealings with opulent Oriental civilizations were bound to produce disparities in wealth, and hence social conflicts, within the aristocracies of Greece. One function of institutions such as guest-friendship was no doubt to ensure the maintenance of the charmed circle of social and economic privilege. This system, however, presupposed a certain stability, whereas the rapid escalation of overseas activity in and after the 8th century was surely disruptive in that it gave a chance, or at least a grievance, to outsiders with the right go-getting skills and motivation. Not that one should imagine concentration of wealth taking place in the form most familiar to the 20th century—namely, coined money. Since 1951 the date of the earliest coinage has been fairly securely fixed at about 600 BC; the crucial discovery was the excavation and scientific examination of the foundation deposit of the Temple of Artemis at Ephesus in Anatolia. The first objects recognizably similar to coined money were found there at levels most scholars (there are a few doubters) accept as securely dated. Coinage did not arrive in Greece proper until well into the 6th century. There were, however, other ways of accumulating precious metals besides collecting it in coined form. Gold and silver can be worked into cups, plates, and vases or just held as bar or bullion. There is no getting round the clear implication of two poems of Solon (early 6th century) that, first, gold and silver were familiar metals and, second, wealth was now in the hands of arrivistes.

The earliest coinage

The first state in which the old aristocratic order began to break up was Corinth. The Bacchiadae had exploited Corinth's geographic position, which was favourable in ways rivaled only by that of the two Euboean cities already discussed. Like Chalcis, which supervised sea traffic between southern Greece and Macedonia but also had close links with Boeotia and Attica, Corinth controlled both a north-south route (the Isthmus of Corinth, in modern times pierced by the Corinth Canal) and an east-west route. This second route was exploited in a special way. Corinth had two ports, Lechaeum to the west on the Gulf of Corinth and Cenchreae to the east on the Saronic Gulf. Between the two seas there was a haulage system, involving a rightly famous engineering feat, the so-called *diolkos*. The *diolkos*, which was excavated in the 1950s, was a line of grooved paving-stones across which goods could be dragged for transshipment (probably not the merchant ships themselves, though there is some evidence that warships, which were lighter, were so moved in emergencies). There is explicit information that the Bacchiadae had profited hugely from the harbour dues. As the Greek world expanded its mental and financial horizons, other Corinthian families grew envious. The result was the first firmly datable and well-authenticated Greek tyranny, or one-man rule by a usurper. This was the tyranny of Cypselus, who was only a partial Bacchiad.

Aristotle, in the 4th century, was to say that tyrannies arise when oligarchies disagree internally, and this analysis makes good sense in the Corinthian context. The evidence of an inscribed Athenian archon list, found in the 1930s and attesting a grandson of Cypselus in the 590s, settled an old debate about the date of Cypselus' coup: it must have happened about 650 (a conclusion for which there is other evidence) rather than at the much later date indicated by an alternative tradition. Cypselus and his son Periander ruled until about 585 BC; Periander's nephew and successor did not last long. Precisely what factor in 650 made possible the success of the partial outsider Cypselus is obscure; no Bacchiad foreign policy failure can be dated earlier than 650. General detestation for the Bacchiadae, however, is clear from an oracle preserved by Herodotus that "predicts" that Cypselus will bring *dikē*, or justice, to Corinth after the rule of the power-monopolizing Bacchiadae. No doubt this oracle was fabricated after the event, but it is interesting as showing that nobody regretted the passing of the Bacchiadae.

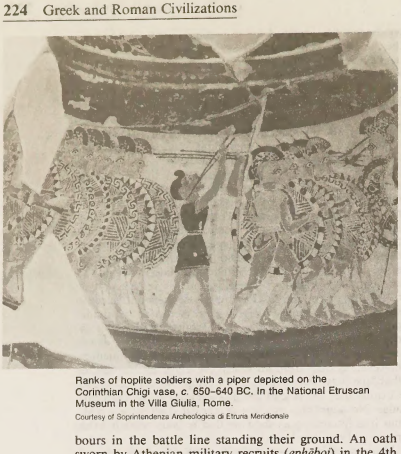
Modern scholars have tried to look for more general factors behind Cypselus' success than a desire in a new world of wealth and opportunity to put an end to Bacchiad oppressiveness and exclusivity. One much-favoured explanation is military, but it must be said straightaway that the specific evidence for support of Cypselus by a newly emergent military class is virtually nonexistent. The background to military change, a change whose reality is undoubted, needs a word.

Aristocratic warfare, as described in the Homeric epics, puts much emphasis on individual prowess. Great warriors used chariots almost as a kind of taxi service to transport themselves to and from the battlefield, where they fought on foot with their social peers. The winner gained absolute power over the person and possessions of the vanquished, including the right to carry out ritual acts of corpse mutilation. This general picture is surely right, though it can be protested that Homer's singling out of individuals may be just literary spotlighting and that the masses played a respectably large part in the fighting described in the epics. There is some force in this objection and in the converse and related objection that in Archaic and Classical hoplite fighting (see below) individual duels were more prevalent than is allowed by scholars anxious to stress the collective character of hoplite combat. Still, a change in methods of fighting undoubtedly occurred in the course of the 7th century.

Change in methods of warfare

The change was to a block system of fighting, in which infantry soldiers equipped with heavy armour, or *hopla* (including helmet, breastplate, greaves, sword, spear, and a round shield attached to the left arm by a strap), fought, at least during part of an engagement, in something like coherent formation, each man's sword arm being guarded by the shield of the man on his right. This last feature produced a consequence commented on by Thucydides—namely, a tendency of the sword bearer to drift to the right in the direction of the protection offered by his neighbour. For this reason the best troops were posted on the far right to act as anchor-men. The system, whose introduction is not commented on by any literary source, is depicted on vases in the course of the 7th century, though it is not possible to say whether it was a sudden technological revolution or something that evolved over decades. The second view seems preferable since the discovery in the 1950s of a fine bronze suit of heavy armour at Argos in a late 8th-century context.

Clearly, the change has social and political implications. Even when one acknowledges some continuation of individual skirmishing, much nonetheless depended on neigh-



Ranks of hoplite soldiers with a piper depicted on the Corinthian Chigi vase, c. 650–640 BC. In the National Etruscan Museum in the Villa Giulia, Rome.
Courtesy of Soprintendenza Archeologica di Etruria Meridionale

bours in the battle line standing their ground. An oath sworn by Athenian military recruits (*ephēboi*) in the 4th century includes clauses about not disgracing the sacred weapons, not deserting comrades, and not handing down a diminished fatherland (to posterity); the oath and the word *ephēbe* are 4th-century, but the institutionalizing of hoplite obligations and expectations is surely much older. Early land warfare can, in fact, be thought of as a symbolic expression of the Greek city's identity. This helps to explain the strong ritual elements in a hoplite battle, which typically began with a sacrifice and taking of omens and ended with victory dedications, often of bronze suits of armour, in some appropriate sanctuary. It is above all the heavily armed troops, not the lightly armed or the sailors in the fleet (nor even the cavalry, whose role in a typical battle was marginal), who were thought of as in a special sense representing the Classical polis. Thus at Classical Athens the 10-tribe citizen system determined the organization of the hoplite army but is much less important in the manning of the fleet.

The influential "hoplite theory" of the origin of tyranny seeks to explain one general phenomenon of the 7th century—namely, the beginning of tyranny—by reference to another, the introduction of hoplite weapons and tactics with their greater emphasis on a collective, corporatist ethos. Insofar as both phenomena represent reactions against aristocratic rule, it is reasonable loosely to associate the two, but it is important to realize that the theory, however seductive, is in its strict form a modern construction. In the first place, the connection is never made by intelligent ancient writers interested both in the mechanics and psychology of hoplite warfare on the one hand and in tyranny on the other. Thucydides, for instance, a military historian if ever there was one, saw tyranny primarily in economic terms. Aristotle does indeed say that the extension of the military base of a state is liable to produce a widening of the political franchise, but this comment has nothing specifically to do with tyranny. He explains tyranny elsewhere either as resulting from splits within oligarchies or by an anachronistic 4th-century reference to demagogic leadership, which, when combined with generalship, is liable to turn into tyranny (here he is surely thinking above all of Dionysius I of Syracuse). In the second place, it is discouraging for the hoplite theory that there is so little support for it in the best-attested case, that of Cypselid Corinth. Attempts have indeed been made to get around the natural implication of the evidence, but they are not convincing. For instance, the ancient statement that Cypselus had no bodyguard ought to be given its natural meaning, which is a denial of the military factor; it ought not to be ingeniously twisted so as to imply that he did not need a bodyguard because (it is argued) he had the support of identifiable army groups. Furthermore, although it is true that Cypselus is called polemarch (which ought to mean a "leader in war"), it is suspicious that his activities in this capacity were entirely civil and judicial. Suspicion increases when one notes that polemarch was indeed the title of a magistrate in Classical Athens.

Other tyrannies are equally resistant to general explanations, except by circularity of reasoning. The Corinthian tyranny has been treated first in the present section because its dates are secure. There is, however, a more shadowy figure, Pheidon of Argos, who may have a claim to be earlier still and who has also been invoked as an exemplification of the military factor in the earliest tyrannies. Unfortunately, one ancient writer, Pausanias, puts him in the 8th century, while Herodotus puts him in the 6th. Most modern scholars emend the text of Pausanias and reidentify Herodotus' Pheidon as the grandson of the great man. This allows them to put Pheidon the tyrant in the 7th century and to associate him with a spectacular Argive defeat of Sparta at Hysiae in 669 BC. His success is then explained as the product of the newly available hoplite method of fighting. (The 8th-century suit of armour from Argos would in fact allow the connection between Pheidon and hoplites even without discarding Pausanias.) This construction assumes much that needs to be proved (see below), and the hoplite theory is in fact being invoked in order to give substance to Pheidon rather than Pheidon lending independent support to the theory.

Pheidon of Argos

Two other tyrannies date securely from the 7th century and perhaps happened in imitation of Cypselus; both arose in states immediately adjoining Corinth. Theagenes of Megara makes an appearance in history for three reasons: he slaughtered the flocks of the rich (an action incomprehensible without more background information than is available); he tried about 630 to help his son-in-law Cylon to power at Athens; and he built a fountain house that can still be seen off the "Road of the Spring-House" in modern Megara. The last two items reveal something interesting about the social and cultural character of established tyrannies (see below), but none of the three offers much support for the military or any other general theory of the cause of tyranny. At Sicyon the Orthagorid tyranny, whose most splendid member was the early 6th-century Cleisthenes, may have exploited the anti-Dorianism already noted as a permanent constituent of the mentality of some Greeks; but since the relevant action—a renaming of tribes—falls in the time of Cleisthenes himself, it is no help with the problem of why the first Sicyonian tyrant came to power at all. In any case, the main object of Cleisthenes' dislike seems to have been not Dorians in general but Argos in particular: the renaming is said to have been done to spite the Argives.

Notwithstanding the skepticism of what has been said above, some solid general points can be made about the tyrannies mentioned (Athens and Sparta followed peculiar paths of development and must be treated separately). First, these tyrannies have more in common than their roughly 7th-century dates: several of the most famous are situated around or near the Isthmus of Corinth. This surely suggests a partly geographic explanation; that is, there was an influx of new and subversive notions alongside the purely material goods that arrived at this central zone. Places with a more stagnant economic and social life, such as Boeotia and Thessaly, neither colonized nor experienced tyrannies. In fact, some version of Thucydides' economic account of the rise of tyranny may be right, though here too (as with the origins of the city-state or the motives behind acts of colonization) one must be prepared to accept that different causes work for different states and to allow for the simple influence of fashion and contagion.

Reflection on the places that avoided tyrannies leads to the second general point. Another way of looking at tyranny is to concentrate on its rarity and seek explanations for that. After all, there were hundreds of Greek states, many of them extremely small, which, as far as is

The rarity of tyranny

known, never had tyrannies. The suggested explanation is that in places with small populations there was enough scope for office holding by most of the city's ambitious men to make it unnecessary for any of them to aspire to a tyranny. (One can add that certain places are known to have taken positive steps to ensure that regular office did not become a stepping-stone to tyranny. For example, a very early constitutional inscription shows that 7th-century Drerus on Crete prohibited tenure of the office of *kosmos*—a local magistracy—until 10 years had elapsed since a man's last tenure.) This is a refreshing approach and surely contains some truth. Nonetheless, the qualification "as far as is known" is important: with regard to many places there is no better reason for saying that they avoided tyranny than for saying that they had it. Moreover, the view that tyranny was widespread may indeed be a misconception, although, if so, it was an ancient one: Thucydides himself says that tyrannies were established in many places. Finally, the psychological argument from satisfaction of ambition is only partly compelling: it was surely more rewarding in every way to be a tyrant than to be a Dreran *kosmos*.

Sparta and Athens. Prominent among the states that never experienced tyranny was Sparta, a fact noted even in antiquity. It was exceptional in this and in many other respects, some of which have already been noted: it sent out few colonies, only to Taras (Tarentum) in the 8th century and—in the prehistoric period—to the Aegean islands of Thera and Melos. It was unfortified and never fully synoecized in the physical sense. And it succeeded, exceptionally among Greek states, in subduing a comparably sized neighbour by force and holding it down for centuries. The neighbour was Messenia, which lost its independence to Sparta in the 8th century and did not regain it until the 360s. It was the Messenian factor above all that determined the peculiar development of Sparta, because it forced Spartans to adjust their institutions to deal with a permanently hostile subject population.

Despite Sparta's military prominence among Greek states, which is the primary fact about it, Sparta's development is especially difficult to trace. This is so partly because there are few Archaic or Classical Spartan inscriptions. Even more important, there is very little genuine Spartan history written by Spartans (there was no Spartan Herodotus or Thucydides, though both men were deeply fascinated by Sparta, as indeed were most Greeks). And partly this is so because—a related point—"invented tradition" had been particularly active at Sparta. As early as the 5th century one finds "laconizers" in other states (the word derives from "Laconia," the name for the Spartan state, or Lacedaemon, and signifies cultural admiration for Sparta and its institutions). The Spartan tradition in European thought can be traced through the centuries up to modern times, though it has never amounted to a single easily definable set of ideas. In the intellectual world of the 4th century BC, when many of the most significant myths about Sparta seem to have been concocted, Sparta, chiefly under the influence of idealist philosophers seeking some solution to civic disorder, was virtually turned into a shorthand expression for a pure community free from *stasis* (internal dissension and fighting) with equality of land ownership and other utopian features that never existed in the historical Sparta or anywhere else. In the Roman period Sparta had become a tourist attraction, a place of uncouth, half-invented rituals. This was also the period when Sparta the living legend consciously traded on and exported fantasies about its great past (in the Hellenistic First Book of Maccabees one even finds the idea seriously put forward that the Jews and Spartans were somehow kin). If more is said about Athens than about Sparta in the present section, that is not because Athens was intrinsically more important but because the amount of usable evidence about it is incomparably greater.

By way of compensation for this lack of evidence about Sparta there are two items of cardinal importance: an extraordinary document about the early Spartan constitution and state, preserved by the Greek writer Plutarch (the "Great Rhetra"), and the poetry of the 7th-century Spartan poet Tyrtaeus. Tyrtaeus wrote poetry in elegiac couplets (alternating hexameter and pentameter lines) intended for symposia. Much of it is military in character and enshrines the hoplite ethic in a developed form at a time when Sparta and Argos were at each other's throats (a fragment of Tyrtaeus' poetry published in 1980 definitely disproved modern skepticism about whether Sparta and Argos could have confronted each other militarily as early as the 660s).

The poetry of Tyrtaeus

Sparta had two kings, or *basileis*. If it is right that this title



Greek expansion (9th–6th centuries BC).

merely denotes hereditary nobles with stated prerogatives, this was originally one of its less remarkable aspects. It is odd, however, that the number two should have been so permanently entrenched. In other respects the Sparta that emerged from the Dark Age had many standard features, such as a warrior assembly based on communal eating in "messes," *syssitia* (a system analogous to the symposium system), and a council of elders. Magistrates called ephors were unique to Sparta and its offshoots, but there is nothing intrinsically odd about formal magistracies.

The Rhetra is an alleged response by the Delphic oracle to the lawgiver Lycurgus around the 9th or 8th century BC. The Rhetra purports to define the powers of the various Spartan groups and individuals just mentioned. It begins, however, by saying that the tribes must be "tribed" (or "retained"; the Greek is a kind of pun) and the obes (a word for a locality) must be "obed." This is desperately obscure, but in an 8th-century context it ought to refer to some kind of political synoecism (Sparta, as stated, was never physically synoecized). The tribes and obes must be the units of civic organization. The Rhetra demands the setting up of a council with the kings and stipulates regular meetings for the Assembly (something not attested at Athens until far later). A crucial final clause seems to say firmly that the people, or *damos,* shall have the power. Yet a rider to the Rhetra, associated with the late 8th-century kings Theopompus and Polydorus, says that, if the people choose crookedly, the elders and kings shall be dissolvers. A poem of Tyrtaeus echoes both parts of this document, rider as well as Rhetra.

The Rhetra is a precocious constitutional document, if it really dates to the 9th or early 8th century, and for this and other reasons (Delphi was not active and writing was not common much before the middle of the 8th century) it is common practice to date the whole document or pair of documents a century or two later. On this view, which is not here followed, the Rhetra itself, with its stipulation of powers for the (hoplite) *damos,* is a 7th-century manifestation of hoplite assertiveness: in fact, it represents a kind of Spartan alternative to tyranny. The references to tribes and obes are then seen as part of a reform of the citizen body and of the army, comparable to and not much earlier than tribal changes elsewhere (see below for Cleisthenes). The rider then dates from an even later period, when Spartan military reverses called for a reactionary readjustment of the balance of power. This view, which involves down-dating Theopompus and Polydorus to the 7th century from the 8th and still more arbitrarily attributing to them the activity presupposed by the Rhetra rather than the rider, does too much violence to the best chronological evidence (that of Thucydides and Herodotus), and a view in terms of 8th-century political synoecism should be preferred. As for the alleged army reform, nothing can be said about it in detail. The best reconstruction is hardly more than a creative fabrication from Hellenistic evidence that dealt with a Spartan religious festival but had nothing straightforwardly to do with the army at all.

It was definitely in the 8th century that Sparta took the step which was to make it unique among Greek states. It had already, in the Dark Age, coerced into semisubject, or "perioikic," status a number of its more immediate neighbours. Then, in the second half of the 8th century, it undertook the wholesale conquest of Messenia (*c.* 735–715). One consequence, already noted, was the export of an unwanted group, the Partheniai, to Taras. These were sons of Spartan mothers and non-Spartan fathers, procreated during the absence in Messenia of the Spartan warrior elite. A still more important consequence of the conquest of Messenia, "good to plow and good to hoe" as Tyrtaeus put it, was the acquisition of a large tract of fertile land and the creation of a permanently servile labour force, the "helots," as the conquered Messenians were now called. The helots were state slaves, held down by force and fear. A 7th-century revolt by the Messenians (the "Second Messenian War") was put down only after decades of fighting and with the help (surely) of the new hoplite tactics. The relationship of hatred and exploitation (the helots handed over half of their produce to Sparta) was the determining feature in Spartan internal life. Spartan warrior peers (*homoioi*) were henceforth subjected to a rigorous military training, the *agoge,* to enable them to deal with the Messenian helots, whose agricultural labours provided the Spartans with the leisure for their military training and life-style—a notoriously vicious circle. The *agoge* and the Sparta that it produced can best be understood comparatively by reference to the kind of male initiation ceremonies and rituals found in other warrior societies. Up to the Second Messenian War, Sparta's political institutions and cultural life had been similar to those in other states. It had an artistic tradition of its own and produced or gave hospitality to such poets as Alcman, Terpander, and Tyrtaeus. But now Spartan institutions received a new, bleak, military orientation. Social sanctions like loss of citizen status were the consequence of cowardice in battle; a system of homosexual pair-bonding maintained the normal hoplite bonds at a level of ferocious intensity; and the economic surplus provided by the lots of land worked by the helots was used to finance the elite institution of the *syssitia,* with loss of full citizen status for men who could not meet their "mess bill." The *agoge,* however, transformed Sparta and set it apart from other states.

Helots

The helot factor affected more than Sparta's internal life. Again and again modifications were forced on Sparta in the sphere of foreign policy. It could not risk frequent military activity far from home, because this would entail leaving behind a large population of discontented helots (who outnumbered Spartans by seven to one). A solution, occasionally tried by adventurous Spartan commanders, was selective enfranchisement of helots. Yet this called for nerve that even the Spartans did not have: on one occasion 2,000 helots, who were promised freedom and were led garlanded round the temples, disappeared, and nobody ever found out what had happened to them. Some person or persons evidently had second thoughts. The Greek historian Xenophon, who was no enemy to Sparta, states that helots (according to the rebel Kinadon) would have liked to eat the Spartans raw, and incidents like this one explain why.

After the suppression of the Messenian revolt (perhaps not before 600), Sparta controlled much of the Peloponnese. In the 6th century it extended that control further, into Arcadia to the north, by diplomatic as well as by purely military means. On the diplomatic level, Sparta, the greatest of the Dorian states, deliberately played the anti-Dorian card in the mid-6th century in an attempt to win more allies. Sparta's Dorianism was unacceptable to some of its still-independent neighbours, whose mythology remembered a time when the Peloponnese had been ruled by Achaean kings such as Atreus, Agamemnon, and his son Orestes (in a period modern scholars would call Mycenaean). The central symbolic act recorded by tradition was the talismanic bringing home to Sparta of the bones of Orestes himself—a way for Sparta to claim that it was the successor of the old line of Atreus. The result was an alliance with Arcadian Tegea, which in turn inaugurated a network of such alliances, to which has been given the modern name of the Peloponnesian League. A valuable 5th-century inscription found in the 1970s concerning a community in Aetolia (north-central Greece) illuminated the obligations imposed by Sparta on its allies: above all, full military reciprocity—*i.e.,* the requirement to defend Sparta when it was attacked, with similar guarantees offered by Sparta in return. Another, more obviously pragmatic, reason why Sparta attracted to itself allies in areas like Arcadia was surely fear of Argos. Archaic and Classical Argos never forgot the great age of Pheidon, and from time to time the Argives tried to reassert a claim to hegemony in mythical terms of their own.

The Peloponnesian League

In the same period (the middle of the 6th century) Sparta drew on its enhanced prestige and popularity in the Peloponnese to take its antipathy to tyranny a stage further: a papyrus fragment of what looks like a lost history supports Plutarch's statement that Sparta systematically deposed tyrants elsewhere in Greece—the tyrannies in Sicyon, Naxos, and perhaps even the Cypselid at Corinth (though this may be a confusion for a similarly named community

called Cerinthus on Euboea). The most famous deposition was Sparta's forcible ending of the tyranny at Athens. Finally one must ask, however, what were Sparta's motives for this and other interventions. Perhaps part of the motive was genuine ideological dislike of tyranny; Sparta was to exploit this role as late as 431, when it entered the great Peloponnesian War as would-be liberator of Greece from the new "tyranny" in Greece—namely, the Athenian empire. Or Sparta may have been worried about the ambitions of Argos, with which certain tyrants, like the Athenian, had close connections. Or it may have longsightedly detected sympathy on the part of certain tyrants toward the growing power of Persia: it is true that Sparta made some kind of diplomatic arrangement with the threatened Lydian power of the Anatolian ruler Croesus not long before his defeat by Persia in 546. If suspicion of Persia was behind the deposition of the tyrants, Sparta was inconsistent in carrying out its anti-Persian policy; it did not help Croesus in his final showdown with Persia, nor did it help anti-Persian elements on Samos (see below), nor did it do much in the years immediately before the great Greek-Persian collision of 480–479 called the Persian War (it sent no help to the general rising of Ionia against Persia in 499 nor to Athens at the preliminary campaign of Marathon in 490). Inconsistency of diplomatic decision making on the part of Sparta is, however, always explicable for a reason already noticed—its helot problem.

Athens was also highly untypical in many respects, though perhaps what is most untypical about it is the relatively large amount of evidence available both about Athens as a city and imperial centre and about Attica, the territory surrounding and controlled by Athens. (This is a difficulty when one attempts to pass judgment on the issue of typicality versus untypicality in ancient and especially Archaic Greek history; it often is not known whether a given phenomenon is frequent or merely frequently attested. This kind of thing creates difficulties for what students of modern history call "exceptionalist" theories about particular states.) Even at Athens there is much that is not yet known; for instance, of the 140 villages, or demes, given a political definition by Cleisthenes in 508, only a handful have been properly excavated.

The distinctiveness of Athens

First, it is safe to say that Attica's huge size and favourable configuration made it unusual by any standards among Greek poleis. Its territory was far larger than that of Corinth or Megara; while Boeotia, though in control of a comparable area, resorted to the federal principle as a way of imposing unity. Like Corinth but unlike Thebes (the greatest city of Classical Boeotia), Athens had a splendid acropolis (citadel) that had its own water supply, a natural advantage making for early political centralization. And Athens was protected by four mountain systems offering a first line of defense. Second, Attica has a very long coastline jutting into the Aegean, a feature that invited it to become a maritime power (one may contrast it with Sparta, whose port of Gythion is far away to the south). This in turn was to compel Athens to import quantities of the ship-building timber it lacked, a major factor in Athenian imperial thinking. (It helps to explain its 5th-century interest in timber-rich Italy, Sicily, and Macedon.) Third, although Attica was rich in certain natural resources, such as precious metal for coinage—the silver of the Laurium mines in the east of Attica—and marble for building, its soil, suitable though it is for olive growing, is thin by comparison with that of Thessaly or Boeotia. This meant that when Athens' territory became more densely populated after the post-Mycenaean depopulation, which had affected all Greece, it had to look for outside sources of grain, and, to secure those sources, it had to act imperialistically. Some scholars have attempted to minimize Athens' dependence on or need for outside sources of grain and to bring down the date at which it began to draw on the granaries of southern Russia via the Black Sea (as it definitely did in the 4th century). Certainly, there were fertile areas of Attica proper, for instance near Marathon, and at many periods Athens directly controlled some politically marginal but economically productive areas such as the Oropus district to the north or the island of Lemnos. A case can also be made for saying that if Athenians had been prepared to eat less wheat and more barley, Athens could have fed itself. Real needs, however, are sometimes less important than perceived needs, and for the understanding of Athenian imperial actions it is more important that its politicians believed (even if modern statisticians would say they were wrong) that internal sources of grain must be endlessly supplemented from abroad. Nor is it entirely plausible to dissociate Athens' 7th-century acquisition of Sigeum (see below) from the provisioning possibilities of the Black Sea region.

Unlike the Peloponnese, with its tradition of Dorian invasion from the north, Athens claimed to be "autochthonous"—that is, its inhabitants had occupied the same land forever. Like any such claim, it was largely fiction, but it helped to make up for Athens' relative poverty in religion and myth: it has nothing to compare with the great legends of Thebes (the Oedipus story) or the Peloponnese (Heracles; the house of Atreus). There was one hero, however, who could be regarded as specially Athenian, and that was Theseus, to whom the original political synoecism of Attica was attributed even by a hardheaded writer like Thucydides.

At whatever date one puts this "Thesean" synoecism, or centralization (perhaps 900 would be safe), it seems that the late Dark Age in Attica saw the opposite process taking place at the physical level; that is, the villages and countryside of Attica were in effect "colonized" from the centre in the course of the 8th century. The process may not have been complete until even later. This explains why Athens was not one of the earliest colonizing powers: the possibility of "internal colonization" within Attica itself was (like Sparta's expansion into Messenia) an insurance against the kind of short-term food shortages that forced such places as Corinth and Thera to siphon off part of their male population.

In fact, Athens did acquire one notable overseas possession as early as 610 BC, the city of Sigeum on the way to the Black Sea. Yet as long as its neighbour Megara controlled Salamis, a large and strategically important island in the Saronic Gulf, the scope for long-distance Athenian naval operations was restricted; the excellent tripartite natural harbour of Piraeus was not safe for use until Salamis was firmly Athenian. Until then, Athens had to make do with the more open and less satisfactory port facilities of Phalerum, roughly in the region of the modern airport. Thus there was an obvious brake on naval expansion.

By the later 7th century then, Athens was looking abroad, and it is not surprising to find it experiencing some of the strains that in the 8th century had led to tyrannies elsewhere. Indeed, it narrowly escaped a first attempt at tyranny itself, that of Cylon, the Olympic victor (630s). Cylon

The close connection between athletic success and military values has been noted; there was an equally close connection between athletic and political achievement, and not just in the Archaic age. Cylon was helped by his father-in-law Theagenes of Megara, a fact that underlines, as does Megarian possession of Salamis until the 6th century, the lateness of Athens' growth to great power status: Classical Megara was a place of small consequence. That Cylon's attempt was a failure is interesting, but too little is known about his potential following to be able to say either that Athenian tyranny was an idea whose time had not yet come or that there is social and economic significance merely in the fact of his having made the attempt.

Cylon's attempt had two consequences for Athenian history. The first is certain but fortuitous: Cylon's followers were put to death in a treacherous and sacrilegious way, which was held to have incriminated his killers, notably Megacles, a member of the Alcmaeonid *genos*. Pollution attracted in this way is a slippery conception; it could wake or sleep, as Aeschylus put it. This particular pollution adhered even to persons who were not on their father's side members of the Alcmaeonid *genos*, such as the great 5th-century leader Pericles, and was usually "woken up" for deliberate and political ends.

The other consequence may not be a consequence at all but a coincidence in time. It was not many years after the Cylon affair that the Athenian lawgiver Draco gave the city its first comprehensive law code (perhaps 621). Because of

the code's extreme harshness, Draco's name has become a synonym for legal savagery. But the code (the purely political features of which are irrecoverably lost to the 20th century short of some lucky inscriptional find) was surely intended to define and so ameliorate conditions; the Athenian equivalents of the "bribe-eating *basileis*" of the Boeotian Hesiod's poem could still dispense a rough, but no longer arbitrary, justice. Further than that it is not safe to go; Draco's code, like that of the statesman and poet Solon (*c.* 630–560), was destroyed by antidemocrats in the late 5th century. A detailed constitution foisted on Draco has survived in the treatise called the *Constitution of Athens*, attributed to Aristotle and found on papyrus in 1890. This says much about the psychology of 411 BC and little about the situation in 621.

Whatever the connection between Cylon and Draco—and one must beware the trap of bringing all the meagre facts about the Archaic period into relation with each other—firmer grounds for postulating economic and social unrest in late 7th-century Attica are to be found in the poetry of Solon. Solon is the first European politician who speaks to the 20th century in a personal voice (Tyrtaeus reflects an ethos and an age). Like the other Archaic poets mentioned, Solon wrote for symposia, and his more frivolous poetry should not be lost sight of in preoccupation with what he wrote in self-justification. He was a man who enjoyed life and wanted to preserve rather than destroy.

Solon's laws

Solon's laws, passed in 594, were an answer to a crisis that has to be reconstructed largely from his response to it. Most scholars believe that Solon's laws continued to be available for consultation in the 5th and 4th centuries; this (as noted above) did not prevent distortion and manipulation. In any case, by the 4th century, the age of treatises like the *Constitution of Athens* and other works by local historians of Attica ("Atthidographers"), much about early Attica had been forgotten or was misunderstood. Above all, there was a crucial failure to understand the dependent status of those who worked on the land of Attica before Solon abolished that status, which was conceived of as a kind of obligation or debt; this abolition, or "shaking off of burdens," was the single most important thing Solon did (see below). When one divides Solon's work, as will be done here for convenience, into economic, political, and social components, one may fail to grasp the possibility that there was a unified vision organizing it all and that in this sense no one reform was paramount. Perhaps the poem of Solon's that sums up best what he stood for is a relatively neglected and not easily elucidated one, but an important one nonetheless, in which he seems to claim that nobody else could have done what he did and still have "kept the cream on the milk." That is to say, his was, in intention at least, a more just though still a stratified society that sought to retain the cooperation of its elite.

Solon canceled all "debt" (as stated, this cannot yet have been debt incurred in a monetary form). He also abolished enslavement for debt, pulling up the boundary markers, or *horoi* (see below), which indicated some sort of obligation. This act of pulling up the *horoi* was a sign that he had "freed the black earth." The men whose land was designated by these *horoi* were called "sixth-parters" (*hektēmoroi*) because they had to hand over one-sixth of their produce to the "few" or "the rich" to whom they were in some sense indebted. Solon's change was retrospective as well as prospective: he brought back people from overseas slavery who no longer spoke the Attic language (this is the evidence, hinted at above, for thinking that the problems facing Solon went back at least a generation, into the period of Draco or even Cylon).

Enslavement for debt was not an everyday occurrence in the world of Aristotle or Plutarch (although the concept never entirely disappeared in antiquity), and they seem to have misunderstood the nature of the debt or obligation that the *horoi* indicated. It is not only Aristotle and Plutarch who found the situation bewildering. It has seemed odd to modern scholars that mere defaulting on a conventional debt should result in loss of personal freedom. Hence they have been driven to the hypothesis that land in Archaic Greece was in a strong sense inalienable and thus not available as security for a loan (of perhaps seed-corn or other goods in kind). Only the person of the "debtor" and members of his family could be put up as a kind of security. Incurable damage has, however, been done to this general theory by the independent dismantling of any idea that land in Archaic Greece was in fact inalienable (such Greek prohibitions on alienation as one hears of tend to date from late and semimythical contexts such as the 4th-century literary reworking of tradition about Sparta or from post-Archaic colonial contexts where the object of equal and indivisible land-portions was precisely to avoid the injustices and agricultural buying-up and asset stripping left behind at home).

Evidently then, some new approach is needed, and it can be found in the plausible idea that what Solon got rid of was something fundamentally different from ordinary debt. In fact, hektemorage was a kind of originally voluntary contractual arrangement whereby the small man gave his labour to the great man of the area, forfeiting a sixth of his produce and symbolically recognizing this subordination by accepting the installation of a *horos* on the land. In return the other perhaps provided physical protection. This would go back historically to the violent and uncertain Dark Age when Attica was being resettled and there was danger from cattle rustlers, pirates (nowhere in Attica is far from the sea), or just greedy neighbours. Alternatively, hektemorage may simply have been the contractual basis on which powerful men assigned land to cultivators in the 9th and 8th centuries, when Attica was being reclaimed after the previous impoverished period. As the 7th century wore on, however, there was scope in Attica for enrichment of an entirely new sort, involving concentration of precious metal in marketable or at least exchangeable form as a result of contacts with elegant, rich, and sophisticated new worlds across the sea. This produced more violent disparities of wealth and a motive for "cashing" the value of a defaulting labourer. On his part, the labourer may have felt that his low social status, once acceptable or inevitable, was no longer commensurate with his military value in the new hoplite age. So Solon's abolition of hektemorage was as much a social and political as an economic change.

This theory of the origin of hektemorage is attractive and explains much. It is disconcerting, however, that the best analogies that can be offered for such semi-contractual "servitude for debt" are from older hierarchical civilizations dependent on highly organized exploitation of man-made irrigation systems (so-called "hydraulic economies"). It is hard to see who or what institution, in Geometric Attica, had the authority—in the absence of any kind of priest-king—to impose the hektemorage system generally throughout the large area of Attica. Nonetheless, one can accept that hektemorage was as much a matter of status as of economic obligation.

Solon's political reforms

Solon's main political changes were first to introduce a Council of 400 members alongside the old "Thesean" council of elders known as the Areopagus, from the Hill of Ares next to the Acropolis, where it met. The functions of this new Council of Solon are uncertain, but that is no reason to doubt its historicity. Solon's Council is perhaps important not so much for itself as for what it anticipated—the replacement Council of Five Hundred, introduced by Cleisthenes at the end of the 6th century.

Second, Solon allowed appeal to the *hēliaia,* or popular law court. The composition of this body is the subject of fierce scholarly dispute; one view sees it as a new and wholly separate body of sworn jurors, enjoying even at this date a kind of sovereignty within the state. The more usual view is that the *hēliaia* was the Assembly in its judicial capacity. This view is preferable: neither in Solon's time nor later is it plausible to posit large juries whose makeup or psychology was distinct from that of the political Assembly. In later times, such appeal to the people was regarded as particularly democratic. But this is just the kind of anachronism one must be careful of when estimating Solon: until pay for juries was introduced in the 460s, such juries could not be a buttress of democracy. Moreover, it would take a courageous peasant (there were no professional lawyers or speech writers as yet) to get up

and articulately denounce a bribe-swallowing *basileus*, especially if—as seems possible—unsuccessful appeal could actually result in increase of sentence.

Third, Solon admitted to the Assembly the lowest economic "class" in the Athenian state, the *thêtes*, whose status was henceforth defined in terms of agricultural produce. The quotes are necessary because investing such fixed economic statuses, or *tele*, with political significance was an innovation of Solon himself; that is, his fourth political reform was to make eligibility for all political office (not just the bare right of attending the Assembly) dependent on wealth and no longer exclusively on birth (a "timocratic" rather than an "aristocratic" system). Solon's four classes were the "five-hundred-bushel men," or *pentakosiomedimnoi*; the *hippeis*, or cavalry class; the *zeugitai*, or hoplites; and the *thêtes*, the class that later provided most of the rowers for the fleet. Again, the immediate impact of the change need not have been cataclysmic: many of the older aristocracy (whether or not one should think of them as a closely defined group of "eupatridae"—that is, "people of good descent") would still have been eligible for office even after the change. But there was also a need to cater to men who were outsiders in the technical sense of not belonging to the older *genē*: the name of one such excluded but high-status category of families has perhaps come down to the present, the so-called *orgeones*. Nor were Solon's four classes themselves entirely new (as indeed the *Constitution of Athens* actually admits in an aside). Thus there were horsemen and even hoplites before Solon, and *thêtes* are mentioned in Homer. The phrase five-hundred-bushel men, which at first sight looks like a prosaic and unimaginative new coinage, acquired in 1968 a 9th-century archaeological analogue: a set of five model granaries was found in a female grave excavated in the Agora. It clearly was a pre-Solonian status symbol ("I was the daughter of a *pentakosiomedimnos*"). An interesting suggestion sees the four classes as originally religious in character: their members may have had allocated functions in the festivals of the synoecized Athenian state. This is not strictly provable but is plausible because the political and military life of Athens and Attica was at all times seen in religious terms.

Solon's social legislation

Solon's social legislation seems generally designed to reduce the primacy of the family and increase that of the community, or polis. To that extent it can be regarded as embryonically democratic. For instance, his laws on inheritance made it easier to leave property away from the family. He also legislated to restrict ostentatious mourning at funerals and to prevent spectacular burials, which were potentially a way for aristocratic families to assert their prestige. (And not just a potential way, either: a great noble called Cimon was buried later in the 6th century in true "Lefkandi style"—that is, close to the horses with which he had won three times at the Olympic games. This burial was surely in defiance of the Solonian rules.) As can be seen from the *Antigone* of the 5th-century tragic poet Sophocles, death and funerary ritual were always an area in which the family, and especially the women, had traditional functions. For the state to seek to regulate them was a major shift of emphasis.

The whole thrust of Solon's reforms was to define and enlarge the sphere of activity of the polis. He was concerned to recognize and increase the power of the ordinary Athenian *thête* and hoplite, while containing without destroying the privileges of the aristocratic "cream." By uprooting the *horoi*, symbols of a kind of slavery, he created as late as the 4th century. And he gave them political rights to match, "as much as was sufficient," as a poem of his puts it. One result of Solon's reforms cannot have been intentional: the abolition of hektemorage created, in modern terms, a "gap in the work force." From then on it was beneath the dignity of the emancipated Athenian to work for a master. Some other source of labour had to be found, and it was found in the form of chattel slaves from outside. That means that the whole edifice of culture and politics rested on the labour of men and women who by "right" of purchase or conquest had become mere things, mere domestic, agricultural, or mining equipment, and whose presence in Classical Attica rose into the tens of thousands. For by the 5th century, slave owning was not confined to the aristocratic few but had been extended to the descendants of that very class Solon had liberated from another kind of slavery.

Initially the Solonian solution was an economic failure, however true it is to attribute to him the economic shape of Classical Attica. Solon himself was almost, but not quite, a tyrant. The orthodox Greek tyrant was associated with redistribution of land and cancellation of debts, though this association was to a large extent a mere matter of popular perception because wholesale redistribution of land is extraordinarily rare in Greek history. Solon did cancel debts. He also redistributed the land in the sense that the former *hektêmoroi* now had control without encumbrance of the land they had previously worked with strings attached. He did not, however, redistribute all the land, because he left the rich in possession of the land the *hektêmoroi* had previously worked for them. In this respect Solon's rule differed from tyranny. It also differed in his simple avoidance of the word; after his year of legislative activity he simply disappeared instead of supervising the implementation of that legislation. This was unfortunate for the former *hektêmoroi*, who needed to be supported in the early years. Growing olive trees, which were a staple of Attica, was an obvious recourse for the farmer in new possession of his own plot, but it takes 20 years for olive trees to reach maturity. Such farmers could hardly look for charity to their former masters, whose wealth and privilege Solon had curtailed. Instead they looked to a real tyrant, Peisistratus.

The Peisistratid tyranny

It took more than one attempt to establish the Peisistratid tyranny, but in its long final phase it lasted from 546 to 510. After the death of Peisistratus, the tyrant's son Hippias ruled from 527 to 510 with the assistance if not co-rule of his brother Hipparchus, who was assassinated in 514.

Hostility to the tyrants on the part of 5th-century informants like Herodotus makes it difficult to ascertain the truth about them. That they ruled with the acquiescence of the great nobles of Attica is suggested by a 5th-century archon list discovered in the 1930s, which shows that even the post-Peisistratid reformer Cleisthenes, a member on his father's side of the Alcmaeonid *genos*, was archon in the 520s. It is also suggested by the fact that Miltiades, a relative of the gorgeously buried Cimon (see above) went out to govern an outpost in the Thracian Chersonese, hardly against the wishes of the tyrants. Furthermore, even the Peisistratids did not confiscate property indiscriminately, though they did levy a tax of 5 percent. This enabled them to redistribute wealth to those who now needed it—that is, those who "had joined him through poverty after having their debts removed (by Solon)." Although a formally ambiguous expression, it must in common sense apply to pre-Solonian debtors, not creditors. How far Peisistratus, who seems to have started as a leader of one geographic faction, specifically mobilized hoplite support at the outset is uncertain, but such military backing is a little more plausible in his century than in the mid-7th century when the "Isthmus tyrants" were seizing power. (Peisistratus' position was, however, buttressed by bodyguards; here, for once, is a tyrant who in some ways fits Aristotle's otherwise excessively 4th-century model.) In any case, Peisistratus' introduction of "demy judges"—that is, judges who traveled round the villages of Attica dispensing something like uniform justice—was an important leveling step, both socially and geographically, and one should imagine this as an appeal to the goodwill of the hoplite and thetic classes. It also, in the longer term, anticipated (as did the well-attested road-building activity of the Peisistratids) the unification of Attica, which Cleisthenes was to carry much further.

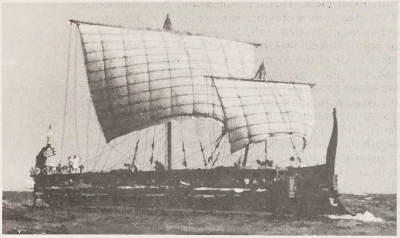
Whether or not Peisistratus climbed to power with hoplite help, he surely strengthened Athens militarily in a way that must have involved hoplites. Indeed, the Peisistratid period ought to count as one of unequivocal military and diplomatic success, and literary suggestions otherwise should be discounted as products of aristocratic malice. In this period should be put the first firm evidence of

the tension between Athens and Sparta that was to determine much of Classical Greek history—namely, Athenian alliances not just with Sparta's enemy Argos but in 519 with Boeotian Plataea. (The Plataeans, faced with coercion from their bigger neighbour Thebes, sued for this alliance at the prompting of Sparta itself; this, however, is evidence of among other things Spartan-Athenian hostility because Sparta's motive, it was said, was to stir up trouble between Thebes and Athens.) Moreover, it may have been in the Peisistratid period that the sanctuary of Eleusis, near Athens' western border and always important for defensive and offensive as well as for purely religious reasons, was fortified. But this is controversial.

This is also the period in which Athens began to be an organized naval power: Salamis became definitively Athenian in the course of the 6th century (tradition credits its taking to both Solon and Peisistratus), with consequences already noted. The island was secured by the installation of what was probably Athens' first cleruchy, a settlement of Athenians with defense functions. Again, it is now that one finds definite mention of the first Athenian triremes, which formed a small private fleet in possession of Miltiades. The trireme, a late Archaic Corinthian invention, was a formidable weapon of war pulled by 170 rowers and carrying 30 other effectives. A full-sized working trireme, launched in Greece in 1987, proved beyond any further debate that triremes were operated by three banks of oars (rather than by three men to an oar). More generally, its size, technological sophistication, and visual impact make it possible to understand Classical Athens' psychological and actual domination of the seas. A proper Peisistratid navy is implied by the tradition that Peisistratus intervened on Naxos and "purified" the small but symbolically important island of Delos, a great Ionian centre. This purification involved ritual cleansing ceremonies and the digging up of graves. As with Eleusis, however, this was deliberate exploitation of religion for the purposes of political assertion.

The trireme

Paul Lipke/Trireme Trust



The Greek trireme *Olympias*.

Elsewhere in Attica also, the Peisistratids interested themselves in organized religion. A literary text first published in 1982 states explicitly what was always probable, that Peisistratus actively supported the local cult of Artemis Brauronia in eastern Attica (the locality from which Peisistratus himself came) and so helped to make it the fully civic cult it is in Aristophanes' play *Lysistrata*. Too much, however, should not be credited to Peisistratus; it has been protested that the relationship between local and city cults in Attica was always one of reciprocity and dialogue. Nevertheless, the explicit evidence about Peisistratus' care for his home cult of Brauron, and the permanent military importance of Eleusis on the way to Peisistratus' enemies in the Peloponnese, make it plausible to suppose a heightening of interest in these two particular sanctuaries precisely in the tyrannical period.

Peisistratid religious and artistic propaganda, and in particular the extent to which the evidence of painted pottery can be used by the political historian, is a modern scholarly battlefield. It has been suggested, on the basis of this sort of evidence, that Peisistratus deliberately identified himself with Heracles, the legendary son of Zeus and Alcmene, and that this is reflected on vase paintings. Yet there are problems; it may be wrong, for reasons already noted, to accord to painted pottery the importance required for the theory. Certainly it needs to be proved that potters, not a numerous or powerful group at any time, had the kind of social standing that would give weight to their "views" as represented on vases, which price lists show were dirt-cheap. In addition, there are particular difficulties about supposing that any man, tyrant or not, could at this early date get away with posing as a god. One is on firmer ground with the Peisistratid building program reflecting not only the tyrant's concern for the water supply (comparable to that shown by the Megarian tyrant) but including the construction of a colossal temple to Olympian Zeus (completed at the time of Hadrian). This and more conjectural buildings on the Acropolis were a direct anticipation of the 5th-century building program of imperial Athens. Unlike painted pottery, they could be commissioned only as a deliberate act by men with plentiful command of money and muscle.

The tyrant Hippias was expelled from Athens by the Spartans in 510. They no doubt hoped to replace him with a more compliant regime, true to their general policy, as described by Thucydides, of supporting oligarchies congenial to themselves. Oligarchy, or rule by the relatively wealthy few, however defined, and tyranny were in 510 the basic alternatives for a Greek state. The newly emancipated Athens of the last decade of the 6th century, however, reacted against its Spartan liberators and added a third member to the list of the political possibilities—democracy. Spartan disappointment at this turn of events expressed itself not merely in unsuccessful armed interventions intended to install a prominent Athenian, Isagoras, as tyrant (506) and even to reinstate Hippias (*c.* 504) but also in attempts to persuade the world, and possibly themselves, that their relations with the Peisistratids had actually been good (hence another source of distortion in the tradition about the tyrants, who on this account emerged as friends, not enemies, of Sparta).

In 508, after a short period of old-fashioned aristocratic party struggles, the Athenian state was comprehensively reformed by Cleisthenes, whom Herodotus calls "the man who introduced the tribes and the democracy," in that order. The order is important. Cleisthenes' basic reform was to reorganize the entire citizen body into 10 new tribes, each of which was to contain elements drawn from the whole of Attica. These tribes, organized initially on nothing more than residence and not on the old four Ionian tribes based purely on descent, would from then on determine whether or not a man was Athenian and so fix his eligibility for military service. The tribes were also the key part of the mechanism for choosing the members of a new political and administrative Council of Five Hundred, whose function it was to prepare business for the Assembly. This Council, or Boule, insofar as it was drawn roughly equally from each tribe, could be said to involve all Attica for the first time in the political process: all 140 villages, or demes, were given a quota of councillors—as many as 22 supplied by one superdeme and as few as 1 or 2 by some tiny ones. An interesting case has been made for saying that this political aspect was secondary and that the Cleisthenic changes were in essence and intention a military reform. Herodotus, for example, remarks on the military effectiveness of the infant Cleisthenic state, which had to deal immediately and successfully with Boeotian and Euboean invasions. And there were arguably attempts, within the Cleisthenic system, to align demes from different *trittyes* (tribal thirds; see below) but the same tribe along the arterial roads leading to the city, perhaps with a view to easy tribal mobilization in the city centre. It is right that the political aspects of Cleisthenes (who was in fact far from producing democracy in the full sense) can too easily be overemphasized at the expense of the military; but the better view is that the new system had advantages on more than one level simultaneously.

The reforms of Cleisthenes

One military result of Cleisthenes' changes is not in dispute: from 501 on, military command was vested in 10 *stratēgoi*, or commanders (the usual translation "generals"

obscures the important point that they were expected to command by sea as well as by land). Normally, each of the 10 tribes supplied one of these generals. They were always directly elected. Direct election for the *stratēgia* remained untouched by the tendency in subsequent decades to move in the general direction of appointment by lot. (Appointment by lot was more democratic than direct election because the outcome was less likely to be the result of manipulation, pressure, or a tendency to "deferential voting.") Even the Athenians were not prepared to sacrifice efficiency to democratic principle in this most crucial of areas. The number 10 remained sacrosanct and so (probably) did the "one tribe, one general" principle, though later in the 5th century, and in the 4th, it was possible for one tribe to supply two generals, one of whom was elected at the expense of the tribe whose candidate had polled the fewest votes. Again, the object was to ensure maximum efficiency: there might be two outstanding men in one tribe. Another peculiarity of the *stratēgia*, to be explained in the same way, was that reelection, or "iteration," was possible. (Actually it is not quite certain that the *stratēgia* was unique in this respect; it is possible that iteration was possible for the archonship as well.)

The *trittys*

The Cleisthenic system was based on the *trittys*, or tribal "third." There were three kinds of *trittys* to each of the 10 tribes, the kinds being called "inland," "coastal," and "city." There were therefore 30 *trittyes* in all, and each of the 140 demes belonged to a *trittys* and a tribe. The numbers of demes in a tribe could and did vary greatly, but the tribes were kept roughly equal in population as far as one can see. (The last words represent an important qualification: it is just possible that the whole system was overhauled in 403 to take account of changes in settlement patterns effected by the great Peloponnesian War. In that case the evidence for deme quotas—evidence which is mostly derived from 4th-century or Hellenistic inscriptions—would not be strictly usable for the 6th or 5th centuries. But in fact there is just enough evidence from the 5th century to make the assumption of continuity plausible.) Each of the 10 tribes supplied 50 councillors to the new Council. In this way, even the remotest deme was involved in what happened in the city; Cleisthenes' solution can thus be seen in its political aspect as an attempt to deal with a characteristic problem of ancient states, which were mostly agriculturally based. That problem was to avoid the domination of city assemblies by the urban population. Cleisthenes' system gave an identity to the deme that it had not had before, even though the word *dēmos* just means "the people," hence "where the people live," hence "village" (the word and concept certainly predate Cleisthenes). Now it had a more precise sense: it was an entity with an identifiable body of demesmen and a right to representation in the Council.

The Cleisthenic deme was the primary unit for virtually all purposes. It was a social unit: to have been introduced to one's demesmen in an appropriate context was good evidence that one was a citizen. It was the primary agricultural unit—though it is disputed whether all settlement in Attica was "nucleated" (that is, whether all farms were clustered together around demes), as one view holds. In fact, there is much evidence for nonnucleated (*i.e.*, isolated) settlement. It was, as stated, a legal unit—although deme judges were suspended from 510 to the 450s. It was a financial unit: temple accounts from the distant deme of Rhamnus date from well back in the 5th century. It was a political unit: as shown, it supplied councillors to the new Council and enjoyed a vigorous deme life of its own (though it seems that there was little overlap between deme careers and city careers). It was a military unit: not only did tribes train together, but a dedication by the demesmen of Rhamnus may show that they participated as a group in the conquest of Lemnos by Miltiades about 500 BC. (Another view puts this inscription in the years 475–450 and sees it as a dedication by cleruchs or a garrison.) Above all it was a religious unit: deme religious calendars, some of the most informative of them published in the 1960s and '70s, show a rich festival life integrated with that of the polis in a careful way so as to avoid overlap of dates. It has been suggested that the worship of Artemis of Brauron, a predominantly female affair, was somehow organized according to the 10-tribe system. Finally, and related to the last, it was a cultural unit: at the deme festival for Dionysus (the "Rural Dionysia") there were dramatic festivals, subsidized, as inscriptions show, by wealthy demesmen.

The Ecclesia

Cleisthenes seems also to have addressed himself to the definition of the Assembly, or Ecclesia. As seen, Solon admitted *thētes* to the Assembly, but Cleisthenes fixed the notional number of eligible Athenians (adult free male Athenians, that is) at 30,000. One-fifth of this total, 6,000, was a quorum for certain important purposes, such as grants of citizenship.

Cleisthenes' ulterior motive in all this must remain obscure in the absence of any corpus of poetry by the man himself, of any biographical tradition, and even of good documentary or historiographic evidence from anywhere near Cleisthenes' own time (the *Constitution of Athens* is reasonably full, but it was written nearly 200 years later).

That the tribal aspect of Cleisthenes' changes was central was recognized even in antiquity, but Herodotus' explanation, that he was imitating his maternal grandfather, Cleisthenes of Sicyon, does not suffice as an explanation on its own. The question is why he should have been anxious that each Athenian tribe should be a kind of microcosm of all Attica. Politically, the tribe does feature in Athenian public life (for instance, tribal support in lawsuits was valuable, and each of the 10 tribes presided by rotation over the Council for one-tenth of the year. This is the so-called prytany system). But the tribe was not a voting unit like the Roman tribe—Athenian votes were recorded as expressions of individual opinion, not submerged in some larger electoral or legislative bloc—and the later political functions of tribes were not quite numerous enough to explain why Cleisthenes felt it necessary to subdivide them into "thirds" in the way he did.

Cleisthenes' changes should be seen in their context. First, the Attica he inherited had a relatively small number of militarily experienced fighters, many of them former Peisistratid mercenaries. It was essential that these be distributed among the tribes if the latter were to be militarily effective. (It is a corollary of this that one accepts that at some preliminary stage in Cleisthenes' reforms there was widespread granting of citizenship to residents of Attica whose status was precarious. There was surely plenty of immigration into prosperous Peisistratid Attica, not all of it military in character.) Second, in the late Archaic period tribal reform took place in other communities, some far removed from Attica in both character and geography. Cleisthenes' system looks subtle, theoretical, and innovatory in its decimal approach to political reform and its reorganization of "civic space," but there were precedents and parallels. For example, at Cyrene, three-quarters of a century after its colonization by Thera, there was *stasis* (political strife), which a reformer, called in from Mantinea on the mainland, settled by reorganizing Cyrene into three tribes. Again, at tyrannical or possibly posttyrannical Corinth, it seems (the evidence is some boundary markers published in 1968) that there was a tribal reorganization along *trittys* lines not dissimilar to, but earlier than, Cleisthenes' system. Finally, there is the Roman analogy: the new system of tribes and centuries, a system based partly on residence, replaced a purely gentilitial system—*i.e.*, one based only on heredity. The word "century" is a clue: although the term signifies a voting unit, it is military in character. It is evident that tribal reform was a fairly general Archaic solution to the difficulties experienced by states with large numbers of immigrants. Such states needed the human resources these immigrants represented, but they could not admit them under the old rules. The rules had to be changed.

Religion and social structure

One may end with religion, which has been called a way of "constructing civic identity" in the ancient world, where religion was something embedded, not distinct. Cleisthenes was a decisive innovator in the social sphere, above all in the new role he allotted to the deme, but he did not dismantle the older social structures with their strong religious resonances. (The phratry, which was associated with Zeus and Apollo, continued to be an important regulator

of citizenship; see above on the Demotionidai inscription.) His 10 new tribes were all named after heroes of Athenian or Salaminian myth, and these tribal heroes were objects of very active cult: this is in itself a recognition of a craving for a religiously defined identity. Nor did the old four Ionian tribes altogether disappear as religious entities; they are mentioned in a sacrificial context in a late 5th-century inscription. The Cleisthenic Athenian state was still in many ways traditional, and it is above all in the religious sphere that one sees continuity even after Cleisthenes.

The world of the tyrants. If the earlier Archaic period was an age of hospitality, the later Archaic age was an age of patronage. Instead of individual or small-scale ventures exploiting relationships of *xenia* (hospitality), there was something like free internationalism. Not that the old *xenia* ties disappeared—on the contrary, they were solidified, above all by the tyrants themselves.

One very characteristic manifestation of this is intermarriage between the great houses of the tyrannical age, as between Cylon of Athens and Theagenes of Megara or between the family of Miltiades and that of Cypselus of Corinth (see above). The Cypselids also were on good terms with the tyranny of Thrasybulus of Miletus in Anatolia (an indication that the Lelantine War alignments had been reversed, though no explanation for this is available). The archetypal event of the Archaic age, however, was the 6th-century entertainment by Cleisthenes of Sicyon of the suitors for the hand of his daughter Agariste. This occasion, to doubt the actuality of which requires the dreariest sort of skepticism, looks back in some respects to the Homeric "suitors" of Penelope in the *Odyssey*. The novelty is that one is now in the world of the polis, and the suitors were men who had "something to be proud of either in their country or in themselves." They came from Italy (two of them, one from Sybaris, one from Siris), Epidamnus in northwestern Greece, Aetolia, Arcadia, Argos (the great-grandson of the great Pheidon), Eretria, Thessaly, and many other places. The winner was one of the two Athenians, Megacles the Alcmaeonid (the other Athenian, Hippocleides, had been well in front but lost the girl by dancing on a table with his legs in the air). Megacles' son by Agariste was the reformer Cleisthenes, named (as so often in Greece) after his grandfather. The suitors were made to perform in the gymnasia (if not too old, Herodotus says), but the decisive "match" at the Trial of the Suitors was held at the final banquet or symposium: proof of the centrality that athletics and communal banqueting had by now assumed.

Anthropological reasons for intermarriage

Although some of the tyrants may (like the Athenian Peisistratids) have retained existing structures such as the archonship and so shown their respect for the status quo, the marriages even of the Peisistratids had politically defiant implications. They were more like pharaonic or Hellenistic sister marriage or like the close intermarrying in aristocratic families of the Roman Republic in that the tyrants had to take their wives only from strains as pure as their own. Yet in the tyrannical world the tyrant had no superiors or equals within his own state. More practically, such ties tended to guarantee political equilibrium. Another related feature that can be explained along similar lines was the practice of multiple marriages (Peisistratus had at least three wives). Breaking the normal social rules in this way had the function of placing the tyrant apart; it is an example of the games princes play.

A third aspect, both cause and consequence, of such intermarriage is internationalism. There also were other factors that contributed to creating something like a common culture or *koinē*. Some of these factors stemmed from an earlier period, such as that of the great Olympic Games (see above *Colonization and city-state formation*). Patronage of poets and artists was a newer phenomenon that helped to make the Greek world a *koinē* through the movement of ideas and individuals from one tyrannical court to another. (The general point must not, however, be exaggerated: cities retained their distinctive cultures, and there were sharp differences of style between one tyrant and another. Even in antiquity the Peisistratids were distinguished from monsters of cruelty like Phalaris, tyrant of Sicilian Acragas.)

The poets Anacreon of Teos and Simonides of Ceos best exemplify the peripatetic life-style of the great cultural figures of the age. Both were brought to Athens by Hipparchus, the son of Peisistratus (Peisistratus himself did not summon poets and musicians to his court, perhaps preferring popular culture like the Dionysia and Panathenaic festivals). Anacreon had previously lived at the court of the splendid Polycrates, the 6th-century tyrant of Samos (who also patronized Ibycus, a native of Rhegium near Sicily); when Polycrates fell, Anacreon was dramatically rescued by Hipparchus, who sent a single fast ship to take him away. Simonides, after the fall of the Peisistratids, moved to the court of the Scopad rulers in Thessaly. Pindar and Bacchylides, the writers of 5th-century victory odes for young aristocrats, were the successors of poets like these. It would be wrong, however, to leave an impression that all the Archaic poets depended on the checkbooks of tyrants; on the contrary, the fragments of Alcaeus of Mytilene on Lesbos (*c.* 600 BC) include invective against the local tyrant Pittacus. And the poetry of his contemporary from the same island, Sappho, has no political content at all but is delicate and personal in character, concerned with themes of love and nature.

More tangible in their achievements, but less easily identified by name, are the tyrannical architects and sculptors, who imitated each other across long distances. The enormous Peisistratid temple of Olympian Zeus is thought to be a direct response to Polycrates' rebuilding of the temple of Hera at Samos; other huge efforts from the same period include a temple at Selinus in Sicily. This frenzied monumentalizing is surely competitive in character, and competition presupposes awareness. Again, Peisistratid interest in the water supply had a parallel not just in the activity of Theagenes at Megara but in a great Polycratean aqueduct at Samos, interestingly, built by a Megarian engineer.

Such eastern Greek influences on thinking in the mainland imply a general Ionian intellectual primacy, which is most obvious in the sphere of speculative thinking. One 6th-century city above all, Miletus in Anatolia, produced a formidable cluster of thinkers (it is best to avoid the metaphor of a series, with its implication that intellectual progress was linear or organized). The cosmological theories of Thales, Anaximander, and Anaximenes are remarkable more for their method—a readiness to work with abstractions, such as water, or the unlimited, to which they accorded explanatory power—than for the actual solutions they reached. It is an interesting modern suggestion that all three were influenced by Persian or even ultimately Indian thought. The suggestion is especially plausible for Heraclitus (fl. 500 BC), because his native city of Ephesus, with its cult of Artemis (a goddess whose worship has features borrowed from that of her native counterpart Anahita) and its large Persian population, was always—down to and including Roman times—especially open to Iranian influences.

Xenophanes

This raises the general question of intellectual awareness of the Persian empire, which conquered the Lydian kingdom of Croesus about 546 BC and so inherited Lydian rule over the Greeks of the Asiatic coastal mainland. The poetry of another poet-philosopher, Xenophanes, from the Ionian city of Colophon, addressed itself to problems of religion and concluded that if horses had gods those gods would be horses, just as Ethiopian gods are black-skinned and Thracian gods have blue eyes. Xenophanes' awareness of the differences between cultures could plausibly be linked to the turnover of empires around him, even if there were no confirmation in the form of a poem describing a symposium at which men "sit drinking sweet wine and chewing chick-pea, and asking each other 'How old were you when the Mede came?' " (The Medes were the predecessors of the Persians, and the Greeks sometimes, as here, conflated the two.) In his "sympotic" aspect—that is, his emphasis on the symposium—Xenophanes was a child of his age; he was more unusual in his rejection, in another poem, of athletic values because of what he thought to be their coarsening effects. One way in which Persia influenced Greek thought was via individual refugees and refugee communities. Thus, Pherecydes of Syros has been seen as a theologian who emigrated from

Anatolia to the west after Cyrus' arrival. (Whether there was a more general westward diaspora of Magi, members of the Persian religious caste, is disputable.) Whole communities left Anatolia under duress; some of them became famous in later philosophical history, such as Phocaea, which founded Elea in Italy, a place famous for philosophy, and Teos, which founded Abdera in northern Greece, the home of the 5th-century atomists Democritus and Leucippus. Finally, one must allow for a considerable Egyptian and western Semitic influence on Archaic Greek religion, political organization, and thought, though its precise extent and the means by which it was mediated await proper scientific treatment.

The greatest literary stimulus provided by neighbouring cultures like the Persian was in the field of ethnography and history. The "inquiries" (*historiai*) of Herodotus, from Asiatic Halicarnassus, will be discussed later, but they would not have been possible without the writings of Hecataeus, another Milesian (*c.* 500 BC), who treated both geography and myth in works that survive today only in fragmentary form. Hecataeus was a "logographer," a prose writer as opposed to the poets so far considered. The gradual move from verse to prose as an intellectual medium goes together with a shift from oral to written culture; but that second shift was not complete even in Athens until well into the 5th century, and there is a case for thinking that even then and in the "document-minded" 4th century "oral" and "written" attitudes coexisted.

Inquiries of Hecataeus' kind had a certain practical application: knowledge of the world, in the most literal sense of that phrase, was of obvious usefulness in a city like Miletus with its colonial connections (in the Black Sea region) and its long-distance trade. (A close connection with Sybaris in southern Italy is implied by Herodotus' story that, when Sybaris was destroyed in 510 BC, the Milesians collectively went into mourning; and Herodotus says that at the beginning of the Ionian revolt, in 500–499, Miletus was at the height of its prosperity.) When the time came to confront Persia politically, after 500, Hecataeus had the standing to suggest initiatives for shared Ionian defense. Surely this standing was conferred as much by what Hecataeus knew as by who he was. On the longer perspective, it was awareness of Persia that helped the Greeks, as it helped the Jews about the same time, to define themselves by opposition. The existence of a great and menacing culture perceived as importantly different was thus a factor in the formation of a common late Archaic Greek culture.

These political and ideological consequences of Archaic Greek thought can be seen as a kind of practical application of theory. The greatest applied scientific achievements of the Archaic period, however, were in the sphere of military technology—the trireme and the hoplite. Some Oriental influence can, it is true, be posited for each (Phoenician for the trireme, Assyrian for hoplite armour); but their refinement and effective use was Greek. The victories of the Persian Wars were won as much by the anonymous Archaic developers of the trireme and the hoplite as by the particular Greeks of 490 and 480–479.

CLASSICAL GREEK CIVILIZATION

The Persian Wars. Between 500 and 386 BC Persia was for the policy-making classes in the largest Greek states a constant preoccupation. (It is not known, however, how far down the social scale this preoccupation extended in reality.) Persia was never less than a subject for artistic and oratorical reference, and sometimes it actually determined foreign policy decisions. The situation for the far more numerous smaller states of mainland Greece was different inasmuch as a distinctive policy of their own toward Persia or anybody else was hardly an option for most of the time. However, Eretria, by now a third-class power, had its own unsuccessful "war" with Persia in 490, and some very small cities and islands were proud to record on the "Serpent Column" (the victory dedication to Apollo at Delphi) their participation on the Greek side in the great war of 480–479. But, even at this exalted moment, choice of sides, Greek or Persian, could be seen, as it was by Herodotus, as having been determined either by preference for local masters or by a desire to spite an equal and rival state next door. (He says this explicitly about Thessaly, which "Medized"—*i.e.*, sided with the Persians—and its neighbour and enemy Phocis, which did not.) Nor is it obvious that for small Greek places the change to control by distant Persia would have made much day-to-day difference, judging from the experience of their kinsmen and counterparts in Anatolia or of the Jews (the other articulate Persian subject nation). Persia had no policy of dismantling the social structures of its subject communities or of driving their religions underground (though it has been held that the Persian king Xerxes tried to impose orthodoxy in a way that compelled some Magi to emigrate). Persia certainly had no motive for destroying the economies of the peoples in its empire. Naturally, it expected the ruling groups or individuals to guarantee payment of tribute and generally deferential behaviour, but then the Athenian and Spartan empires expected the same of their dependents. The Athenians, at least, were strikingly realistic and undogmatic about not demanding regimes that resembled their own democracy in more than the name.

The Ionian revolt

But the experience of the Asiatic Greek cities was different again, because it was precisely here that the great confrontation between Greeks and Persians began, about 500 BC. The first phase of that confrontation was the "Ionian revolt" of the Asiatic Greeks against Persia (despite the word Ionian, other Asiatic Greeks joined in, from the Dorian cities to the south and from the so-called Aeolian cities to the north, and the Carians, not Greeks in the full sense at all, fought among the bravest). The puzzle is to explain why the revolt happened when it did, after nearly half a century of rule by the Achaemenid Persian kings (that is, since 546 when Cyrus the Great conquered them; his main successors were Cambyses [530–522], Darius I [522–486], Xerxes I [486–465], Artaxerxes I [465–424], and Darius II [423–404]). Too little is known about the details of Persian rule in Anatolia during the period 546–500 to say definitely that it was not oppressive, but, as stated above, Miletus, the centre of the revolt, was flourishing in 500.

The causes of the Ionian revolt are especially hard to determine because the revolt was a short-term failure. (Concessions were made after it, however, and its longer-term consequence, the Persian Wars proper, resulted in the establishment of a strong Athenian influence in western Anatolia alongside the Persian.) Defeats lead, especially in oral traditions, to recriminations: "Charges are brought on all sides," Herodotus says despairingly about the difficulty of finding out the truth about the crucial naval battle of Lade (495). Herodotus himself was contemptuously hostile, regarding the revolt as the "beginning of troubles"—a phrase with a Homeric nuance—between Greeks and Persians. This is odd, because it is inconsistent with the whole thrust of his narrative, which regards the clash as an inevitability from a much earlier date; it is part of his general view that military monarchies like the Persian expand necessarily (hence his earlier inclusion of material about, for instance, Babylonia, Egypt, and Scythia, places previously attacked by Persia). The reasons for Herodotus' hostility have partly to do with anti-Milesian sentiment specifically in fellow-Ionian Samos, where he gathered some of his material (the Samians seem to have tried to represent the failure as due to the incompetence and ambitions of Milesian individuals), and partly with the generally Ionian character of the revolt (Herodotus' home town of Halicarnassus was partly Dorian, partly Carian). In addition, he was influenced by defeatist mainland Greek sources, particularly by Athenian informants who resented Athens' unsuccessful involvement on the rebel side. And he genuinely thought that the Persian-Greek conflict was a horrible thing, although mitigated, in his view, by the fact that Persians and Greeks, particularly Spartans, gradually came to know each other and respect each other's values. There were always Greeks who were attracted to a Persian life-style.

It should now be clear that Herodotus saw the revolt in terms of the ambitions of individuals (he singles out the Milesians Aristagoras and Histiaeus), and this must be

part of the truth. But this must be supplemented by deeper explanations, because the rising was a very general affair.

Causes of the Persian Wars

A simple economic explanation, such as used to be fashionable, is no longer acceptable. Perhaps one should look instead for military causes: Ionians disliked the military service to which they were then compelled (they did not even care much for the naval training they had to undergo, in a better cause, before Lade). Persia not only expected personal military service but punished attempts to evade it, even at high social levels. Its method of organizing defense and of raising occasional large armies (there was no large Persian standing army) was analogous to the method of later feudalism: "fiefs" of land were granted in exchange for political loyalty and for military service when occasion required.

Here perhaps is a clue, which permits the resurrection of the economic explanation in another more sophisticated form. Grants of fiefs in Anatolia are well attested in the 5th and 4th centuries; in the pages of the Greek historian Xenophon (431–350) one finds the descendants of Medizing Greek families still installed on estates granted to their ancestors after 479 (and inscriptions show the same families were still there well into the Hellenistic period). Grants by Persia of good western Anatolian land to politically amenable Greeks, or to Iranians, made good political and military sense. Such gifts, however, were necessarily made at the expense of the poleis in whose territory the land so gifted had lain. In this, surely, were the makings of a serious economic grievance.

Politically, the Greeks did not like satrapal control. This seems clear from the concessions made after the revolt ended in 494: the Persians Artaphernes and Mardonius granted a degree of autonomy by instituting a system of intercity arbitration; they abstained from financial reprisals and from demanding indemnities and merely exacted former levels of tribute, but after a more precise survey; and above all, Herodotus says, they "put down all the despots throughout Ionia, and in lieu of them established democracies." The meaning and even the truth of this last concession are alike disputed. Although there certainly were still tyrants in some Persian-held eastern Greek states in 480, some improvement on arbitrary one-man government is surely implied. Perhaps the answer is to be found in the formula recorded by a later literary source, the Greek historian Diodorus Siculus (fl. 1st century BC), who wrote that "they gave them back their laws." (When in 334 Alexander similarly claimed to restore to the Ionian and Aeolian cities their laws and democracies, he was largely indulging in propaganda.) Inscriptions, above all from Persian-occupied Anatolia in the 4th century, show that the cities in question held tribal meetings, enjoyed a measure of control over their own citizen intake, levied city taxes (subject to Persia's overriding tribute demands), and did indeed operate a system of intercity arbitration. How different all this was from the situation before 500 is beyond retrieval, but the continuity of civic structures and cults in eastern Greek states from the Archaic period to Classical times implies that in many respects the Persian takeover of 546 was not cataclysmic. For instance, one reads at the very end of Herodotus' history (concerning the year 479) of a temple on Asian soil to Demeter of Eleusis that had been brought over by the Ionians from Attica in the early Dark Age and was still going strong, presumably without a break. So the improvements introduced after 494 consisted in the increase, not in the outright introduction, of local self-determination within the satrapal framework.

In any case, one is left with the problem of why political unrest boiled over, if boil over it did, in precisely 500. A large part of the answer is to be found in the changes recently made at the Ionian mother city Athens by Cleisthenes. Local arrangements that may have seemed tolerable before the end of the century seemed less so in face of the new political order at Athens, an order that had moreover shown its military effectiveness. The hypothesis that the example of Cleisthenic Athens induced restlessness elsewhere is plausible not just for its kinsmen in Ionia, which can be supposed to have had good "colonial" communications with Athens, but even for the Peloponnese, where in the first half of the 5th century Sparta had to deal with persistent disaffection (see below).

Athenian support of Ionia

Communication between Athens and Ionia in this period is, however, first firmly attested in the other direction, not to Ionia but from it. In 499 the Milesian tyrant Aristagoras arrived in Athens and Sparta (and perhaps at other places too, such as Argos) asking for help. The Athenians agreed, while the Spartans under their king Cleomenes (who ruled from 519 to shortly before 490) did not, thus showing, as Herodotus says, that "it seems indeed to be easier to deceive a multitude than one man." This is out of line with Herodotus' otherwise favourable assessment of Cleisthenic democracy and should be put down to particular hostility to the revolt and its consequences for Athens. The Athenians sent 20 ships. This was a major undertaking, considering Athens' resources and commitments; in 489 (when Athens' fleet was surely bigger than it had been a decade earlier) Athens had only 70 ships, of which 20 were borrowed from Corinth. The reason Athens had borrowed these ships from Corinth (actually it was a sale at nominal charge) was Athens' war, or series of wars, with Aegina, which had caused it to build a fleet. Corinth and Athens, both of which had naval outlets in the Saronic Gulf, had a shared interest in containing the power of Aegina, the greatest other power in that gulf, the "star in the Dorian Sea," as Pindar was to call Aegina. The Athenian-Aeginetan struggle, which may actually have continued after the Battle of Salamis in 480, having begun well back in the late 6th century with a shadowy precursor in the mythical period, meant that the Athenian help sent to Ionia was risky and heroic.

On a longer perspective the struggle against Aegina helped to make Athens a naval power through simple peer-polity pressure. Ancient versions of the Athenian ship-building program, however, put too much onto the Aeginetan factor, usually out of malice against the great Athenian politician Themistocles and reluctance to give him credit for anticipating the eventual arrival of the Persian armada of 480. The better tradition allows Themistocles an archonship in 493, during which he started the walls of the Piraeus, turning it into a defensible harbour, and so first "dared to say that the Athenians must make the sea their domain" (as Thucydides puts it with forgivable exaggeration). The Ionian revolt had failed disastrously, Miletus having been sacked in 494, and it was clear that the Persian finger was now pointed at Athens and that Darius wanted revenge for the assistance it had sent. The result was the Marathon campaign.

Sparta did not participate in the Battle of Marathon. Spartan policy toward Persia in particular, and its foreign policy in general in the years 546–490, is at first sight indecisive. Having expelled the pro-Persian Peisistratids, Sparta not only tried to put them back a few years later but declined to help the Ionians in 499. The reason given by Cleomenes on that occasion, after a glance at the position of the Persian capital Susa on the map, was that it was outrageous to ask a Spartan army to go three months' journey from the sea. This is a colourful way of saying that it was a tall order to ask Sparta to go to the help of distant Greeks, with few of whom it had kinship ties. The Spartans who had made the original, admittedly ineffective, alliance with Croesus against the Persians had not been so timid. But that was when the Persian threat had scarcely appeared over the horizon. In 490 the reason for Spartan nonappearance at Marathon was a religious scruple: the Spartans had to wait until the Moon was full, probably because this was the sacred month of a festival. There is no good reason to doubt this, though it has been argued that there were special reasons why Sparta's leadership was halfhearted in the 490s and that it should be related to the scattered evidence for helot trouble at precisely this time.

Cleomenes' own career ended in disgrace not long before Marathon. It has been suggested that he fell foul of the domestic authorities at Sparta (who always had the power to discipline the kings) because he made promises to the helots: he proposed freedom in exchange for military service. If so, this must have been late in his career; the reply to Aristagoras in 499 looks straightforward. In any case,

the theory rests largely on the equally speculative theory that replaces the religious explanation for Spartan absence from Marathon with a political one. To say this is not to deny the permanent threat posed by the helots, still less to deny that Spartan equivocation can often be explained in terms of it.

Cleomenes

Large claims have been made for the statesmanship of Cleomenes, but his vision does not seem to have gone beyond the narrow issue of "What is best for Sparta?" For instance, Cleomenes crushed the old enemy Argos, then resurgent, at the great battle of Sepeia (near Tiryns) in 494. He was too shrewd, however, to destroy it completely, realizing that dislike of Argos was one of the factors that kept Sparta's Peloponnesian allies loyal. Argos was left to the control of a group described as "slaves" (hardly literally that, perhaps really members of surrounding subject communities), which was thoroughly traditional Spartan behaviour on Cleomenes' part. He surely does not deserve to rank as a forward-looking "Panhellenist"—that is, as a supra-Spartan enemy of Persia. While it is true that he did act on one occasion against Medizers on Aegina, he did so only at the 11th hour, perhaps as late as 491. Even by the criterion of Sparta's local interests, Cleomenes, or more fairly Sparta's treatment of Cleomenes, had bad results. Cleomenes' offer of some kind of new deal to the Arcadians (better substantiated than his dealings with the helots) came to nothing with his spectacular death; he went insane (it was alleged), was imprisoned, and committed suicide. Some Arcadian states were certainly disaffected in the 470s and 460s and perhaps even anticipated 4th-century developments by forming a (numismatically attested) league of their own. It is also tempting to link the new pattern of forces in the Peloponnese, which enabled Argos to recover sufficiently to conquer Mycenae (460s) while Sparta was preoccupied elsewhere, with the activities of Cleomenes toward the end of his life and the expectations he had aroused only to disappoint. (Another plausible factor in Arcadia then, as in Ionia in 500 BC, was the unsettling effect of Cleisthenic democracy at Athens.) At least one can say that Spartan worries about Arcadia were relevant to the "Great Refusal" of leadership in 479, which made possible the Athenian empire.

Athens was not entirely alone in its fight against the Persians at the Battle of Marathon in 490 BC. Plataea fought beside Athens, true to the alliance of 519, and the Tomb of the Plataeans, excavated in 1966, probably commemorates the place where they fell. Eretria, which had also sent help to the Ionian revolt, had already been pounced on and destroyed. The reasons for the Persian choice of Marathon, as given by Herodotus, were proximity to Eretria (that is, the Persians wanted a short line of communications) and the good cavalry terrain there. He does not add, however, that a third powerful motive was political. The deposed Peisistratid tyrant Hippias, now a bitter old man, accompanied the Persian forces. (The Peisistratids came originally from eastern Attica.) Cleisthenes, in implementing his democratic reforms after the fall of the tyrants, had perhaps tried to break up old sources of political influence in this region. For instance, Rhamnus, a little to the north of Marathon and a vital coastal garrison site in Classical and Hellenistic times, seems to have been anomalously attributed to a city *trittys;* and an ancient local organization known as the Marathonian "Four Cities," or Tetrapolis, was broken up among more than one of the new tribes. Reasonably or unreasonably, Hippias was obviously hoping to establish a kind of political bridgehead here by appealing to old bonds of clientship.

The Battle of Marathon

The Athenians, however, marched out immediately under Miltiades, who had been recalled a year or two earlier from the Chersonese to help Athens meet the danger. Then, perhaps when the Persian cavalry was temporarily absent, they attacked the Persians "at a run." This last detail impressed itself on the tradition, as it undoubtedly impressed the waiting Persians, and the discovery of Persian arrowheads in the Athenian burial mound makes it possible to supply the explanation that had eluded Herodotus. The Athenian advance was evidently achieved under a hail of arrows; and the quicker the dangerous ground was covered, the better.

The Athenian victory was overwhelming; there were 6,400 Persian casualties to 192 Athenian. It was an important victory for two reasons. First, it showed what lethal damage hoplites could do to Persian forces; this encouraging message was not missed by the Spartans who arrived to view the corpses and departed with patronizing congratulations to the Athenians. Second and more important, it was a propaganda victory, celebrated in all the available media. Marathon soon became an almost mythical event. The Athenian Treasury at Delphi was built out of the spoils of the battle. An ambitious conjecture seeks to equate the 192 Marathon dead with the 192 equestrian figures on the Parthenon frieze. The horses on the frieze would be a difficulty if the idea was to recall the battle in a literal way, because the battle was definitely not a cavalry affair; but it has been ingeniously suggested that the horses were intended to suggest "heroic" status in the technical sense of "hero," or demigod. Heroic cult often involved horses (as perhaps at Lefkandi), and heroic funerals regularly included equestrian events. This interpretation, however, poses problems for two reasons: the frieze was partially destroyed in the 17th century and reconstruction depends on old drawings, and the evidence for actual heroization of the Marathonian dead is late.

Still, there is no doubting the symbolic significance of Marathon, or the way in which well after the Persian Wars the victory was exploited in epigram and painting. For instance, there was a famous rendition of the Battle of Marathon in the "Painted Colonnade" at Athens (now lost), which was perhaps commissioned by Miltiades' son Cimon. This was celebratory artistic propaganda, with a far clearer message than that of the Peisitratids. The Battle of Marathon and the Persian Wars must be recognized as an artistic watershed. There was admittedly something splendid about the gesture of sending help to the Ionian revolt, and it has been suggested accordingly that early 5th-century depictions on vases of Theseus attacking the Amazons (inhabitants of Anatolia) may be a coded allusion to Athens' Asiatic adventure of the 490s. The vindictive Athenian treatment of the playwright Phrynichus for referring in a play to the fall of Miletus shows, however, that the Ionian revolt was a dangerous subject, not lightly to be treated by pot painters. Marathon was the beginning of an epoch that lasted for centuries, during which Athens asserted its claim to uniqueness on the basis of two things: its achievements in the Persian Wars and (in and after the 4th century) its cultural primacy.

Meanwhile the Persians retreated, and Darius died, to be succeeded in 486 by Xerxes. No Greek could have doubted that Marathon, for all its symbolic importance, was not the end of the matter and that Xerxes would return with a much larger invasion force.

Appointment of archons

The internal Athenian reaction to this latest military success of the Cleisthenic democracy was to take the development of that democracy a stage further. First, a change was made in the method of appointment to the chief magistracy, the archonship. From then on the archons were appointed by lot from a preliminary elected list instead of being directly elected, as the *stratēgoi* continued to be. There were nine archons and a secretary. Three of the archons had special functions: the *basileus,* or "king"; the *polemarchos,* originally a military commander (though after the institution of the Cleisthenic *stratēgoi,* military authority passed to this new panel of 10); and the "eponymous archon," who gave his name to the year. Interpretation of the significance of the change varies according to the view taken of the importance of the archonship itself in the period 508–487; perhaps it was a young man's office and of no great consequence. The period is patchily documented, however, and in any case it would be eccentric to query the distinction of some of the names preserved. The point has a bearing on the composition and authority of the ancient Council of the Areopagus, which was recruited from former archons. The role of the Areopagus was to be much reduced in the late 460s, and if the archonship was after all not especially prestigious, then the importance of that subsequent attack on the Areopagus would be correspondingly reduced. A more substantial reason for thinking that the archonship

mattered less after 508 than it had, for instance, under the Peisistratids lies in the "seesaw" argument that the rise of the *stratēgia* must have led to a fall in the power and prestige of the archonship.

The reform of 487 was probably the first time that lot or "sortition" had been used, though it is possible that Cleisthenes, or even Solon, used it as a device for distributing posts equitably among basically elected magistrates. This would not be unthinkable in the 6th century, when the Athenian state still contained so many aristocratic features; after all, the Romans used sortition in this way, not as a consciously "democratic" procedure but as a way of resolving the competing claims of ambitious individuals.

There is a further slight uncertainty about the system of "sortition from an elected shortlist." The usual and probably correct view is that this system was discarded, not long after 457, for the archonship and other offices appointed to by lot in favour of unqualified sortition. But there is enough evidence for the survival of the preliminary stage of election to have encouraged a theory that the hybrid system continued in use down to the 4th century. This, if true, would have serious implications for our picture of Athenian democracy, but the best evidence for the hybrid system is in untypically conservative contexts, such as appointment to deme priesthoods.

A further novelty of the early 480s was the first ostracism. This was a way of getting rid of a man for 10 years without depriving him of his property. First, a vote was taken as to whether an ostracism should be held in principle; if the voters wanted one, a second vote was taken, and, if the total number of votes now cast exceeded 6,000, the "candidate" whose name appeared on the largest number of potsherds, or ostraca, went into this special sort of exile. An obstinate tradition associates the introduction of ostracism with Cleisthenes, but this hardly matters because the evidence is explicit that no ostracism was actually held until 487. The object of this very unusual political weapon has been much discussed; whereas some ancient writers considered it as a way of preventing a revival of the Peisistratid tyranny (hardly a real threat after 490), modern scholars see it as a device for settling policy disputes—that is, as a kind of ad hominem referendum. It is possible, however, to be too rational about ostracism; Ostracism of the large numbers of ostraca that survive, not all have been completely published, but one can see that their content is sometimes abusive and sometimes obscene. One accuses Cimon of incest with his sister, another says that Pericles' father Xanthippus "does most wrong of all the polluted leaders." The idea of the politician-leader as polluting scapegoat is a recurrent one in Greek political invective, and it is perhaps in terms of invective, or the need for a religious safety valve, that ostracism can best be understood.

Evidently, the Athenian demos was growing more bold, as the *Constitution of Athens* puts it. This was equally true in foreign affairs. The year after Marathon, Miltiades made an attack on the Aegean island of Paros, which anticipates the more systematic imperialism of the period after 479. And it is possible that the Athenian duel with Aegina continued into the 480s. But the event with the greatest implications for foreign policy was a sudden large increase in the output of the Laurium silver mines (actually in a region called Maronea). The evidence gives the crucial year as 483, but it is not known whether there was really a dramatic lucky strike just before that or whether this was merely the year when Athens decided how to spend the accumulated yield of several good years. One source does speak of "discovery" of mines, but the mining area had been worked since Mycenaean times, and the mines were certainly operational under the Peisistratids. It was decided to spend the windfall on building more triremes, bringing the total to 200 by 480, from the 70 attested for Miltiades' Parian expedition of 489. The precise method somehow involved the advancing of money to individuals, an interesting partial anticipation of the Classical system of "trierarchies." Trierarchs, who are not specifically attested before the middle of the century, were wealthy individuals who, as a kind of prestige-conferring tax payment, paid for the equipping of a trireme (the state provided the hull). The source of the timber for this huge program is not known; perhaps local Attic or Euboean supplies supplemented Italian timber. Themistocles, who is credited with the essential decision to spend the money on ships rather than on a distribution among the citizens, had western interests that make the Italian hypothesis plausible. If this is right, the feat of transportation should be admired almost as much as the crash building program itself. One consequence of the rapidity of the program was that much of the timber must have been unseasoned; this is relevant to the Greek decision at the Battle of Salamis (480) to fight in narrow waters, where the resulting loss of speed (green timber makes ships heavier and slower) would matter less.

It was in Athens, then, that the most energetic action was taken. Xerxes had not lost sight of the old revenge motive, a motive that ought to have meant that Athens was the main or only target, but his aim this time was—as Herodotus correctly says—to turn Greece as a whole into another Persian province or satrapy. This called for a concerted Greek plan, and in 481 the key decisions were taken by a general Greek league formed against Persia. Quarrels like that between Athens and Aegina had to be set aside and help sought from distant or colonial Greeks such as the Cretans, Syracusans, and Corcyrans, whose extraordinarily large fleet of 60 ships (possibly developed against Adriatic piracy but also—surely—against Corinth) would be a prime asset. Corcyra, however, waited on events, and Crete stayed out altogether, while Syracuse and Sicily generally had barbarian enemies of their own to cope with, the Carthaginians. (Syracuse and other parts of Sicily were now under the tyranny of Gelon.) Greek writers found the parallel between the simultaneous threats to eastern and western Hellenism irresistible and represented Carthage as another Persia. It has, however, been suggested that the imperialistic ambitions of Carthage have been generally exaggerated by Greek writers eager to flatter their patrons, such as Gelon. The reality of the Battle of Himera, however, in which Gelon decisively defeated the Carthaginians, is not in doubt; like the Battle of Salamis, it was fought in 480, allegedly on the same day. Gelon did indeed have his own preoccupations. The Greeks may not have been altogether sorry: the tyrant Gelon would have been an ideologically awkward ally in a struggle for Greek freedom from arbitrary one-man rule. Gelon and Carthage

Even without western Greek help, the Greek fleet numbered about 350 vessels, amounting perhaps to a third of the Persian fleet. The size of the Persian land army is reckoned in millions by Herodotus, and all modern scholars can do is replace his guess by far lower ones.

Greek unity, though impressive, was not complete; conspicuous among the "Medizers" was Thebes, while Argos' neutrality amounted, in Herodotus' view, to Medism.

An inscription found in 1959, the so-called "Decree of Themistocles," purports to contain further detailed decisions made about this time regarding the evacuation of Attica and the mobilization of the fleet. But the writing is of the 4th century, and the whole text is probably not a re-inscribing of a genuine document but a patriotic concoction of the age in which it was written and erected.

An initial plan to defend Thessaly was soon abandoned as unrealistic. Instead the Greeks fell back on a zone at the northeastern end of Euboea, where they hoped to defend Thermopylae by land and Artemisium by sea. Herodotus, who is often accused of failing to realize the interconnectedness of these two holding operations, did in fact stress that the two were close enough for each set of defenders to know what was happening to the other.

The Spartans had sent their king Leonidas to Thermopylae with a force of 4,000 Peloponnesians, including 300 full Spartan citizens and perhaps a helot contingent as well. They were joined by some central Greeks, including Boeotians from Thespiae and Thebes. The pass at Thermopylae could not be held indefinitely, as Leonidas surely knew, but he also knew that an oracle had said that Sparta would be devastated unless one of its kings was killed. Leonidas' exact "strategy" has been debated as if it were a puzzle, but perhaps one should not go beyond the oracle. The king must die.

Leonidas died, with his 300 Spartans (and the helots, Thespians, and Thebans, as should be remembered to the honour of all three). The other groups, Peloponnesians and central Greeks, were all dismissed. The naval action at Artemisium was inconclusive, the real damage to the Persian ships being done by a storm as they rounded Euboea.

Whether or not the Decree of Themistocles is genuine, it is a fact that Attica was evacuated and the Athenian Acropolis sacked by the Persians. This sacrifice of their city, like the victory of Marathon, is one of the cardinal elements in Athenian celebration of the Persian Wars. The Persians destroyed the temples on the Acropolis and carried off the statues of Harmodius and Aristogiton, the two men who had assassinated the tyrant Hippias' brother Hipparchus in 514. The symbolic importance to Athens of what happened on the Acropolis in 480 is illustrated by the subsequent history of those statues: they were returned to Athens by Alexander the Great a century and a half later as part of his propaganda claim to be punishing the Persians for their 5th-century impiety.

The Battle of Salamis

The Persians entered the narrows of Salamis, where Themistocles had insisted the Greeks should be stationed, and they were comprehensively defeated under the appalled eyes of Xerxes himself. This defeat is a "David and Goliath" encounter only in the general sense that the Persian empire was vastly greater, in size and resources, than the realm of its Greek opponents. It is said that the Greek ships were actually heavier and less easy to maneuver than those of their opponents. Yet this Persian advantage, and that conferred by the greater experience of the Phoenician sailors on the Persian side, were canceled out by the Greek advantages of position: a fight in the narrows would enable them to board and fight hand to hand. No doubt there was also a propaganda aspect. Themistocles had inscribed on the rocks of Euboea messages imploring the Ionians on the Persian side not to help in enslaving their Ionian kin. This looks forward to Athens' political exploitation, in the very near future, of its original role as Ionian mother city. For the moment it surely helped sap morale in the Persian fleet. Sound strategy might have dictated a Persian withdrawal, or an attempt to bypass Salamis and press on to the Isthmus of Corinth, before the battle had even begun, but the prestige of the Persian king was visibly at stake.

Xerxes returned home, but the Persian general Mardonius remained for a final encounter with the Greeks at Plataea. The Spartans under Pausanias, regent for the underage Spartan king, advanced from the Peloponnese via the Isthmus and Eleusis; there had once been a question of making a stand at the Isthmus for the defense of the Peloponnese, but Salamis had made that unnecessary. Again the Persians were defeated, but this time the battle was primarily won, as Aeschylus was to put it in his play *Persians*, "under the Dorian spear"—that is, under the leadership of hoplite Sparta. (The army, however, was a truly Pan-Greek one and included a large infantry force of Athenians.) As much glory was to attach to Plataea itself as to Sparta. A Hellenistic geographer said with some impatience of the Plataeans that they had nothing to say for themselves except that they were colonists of the Athenians (strictly false, but an illuminating exaggeration) and that the Persians were defeated on Plataean soil. A great commemorative festival was still celebrated at Plataea in Hellenistic and Roman times; a 3rd-century inscription discovered in 1971 mentions "the sacrifice in honour of Zeus the Liberator and the contest which the Greeks celebrate on the tombs of the heroes who fought against the barbarians for the liberty of the Greeks." After the residue of the Persian fleet had been defeated at Mycale, on the eastern side of the Aegean, the Greeks were saved—for the moment. The Persians had, after all, returned to Greece after the small-scale humiliation of Marathon in 490; thus there could be no immediate certainty that they would abandon their plans to conquer Greece after the far greater humiliations of 480 and 479. A leader was required in case the Persians returned.

The Athenian empire. The eastern Greeks of the islands and mainland felt themselves particularly vulnerable and appealed to the natural leader, Sparta. The Spartans' proposed solution was an unacceptable plan to evacuate Ionia and resettle its Greek inhabitants elsewhere; this would have been a remarkable usurpation of Athens' colonial or pseudocolonial role as well as a traumatic upheaval for the victims. Samos, Chios, Lesbos, and other islanders were received into the Greek alliance. The status of the mainlanders was temporarily left in suspense, though not for long: in early 478 Athens on its own account captured Sestus, still under precarious Persian control hitherto. In this it was assisted by "allies from Ionia and the Hellespont"—that is to say, including mainlanders. The authority for this statement, which should not be doubted, is Thucydides, the main guide for most of the next 70 years.

The capture of Sestus was one manifestation of Athenian independence from Spartan leadership, which had gone unquestioned by Athens in the Persian Wars of 480–479, except for one or two uneasy moments when it had seemed that Sparta was reluctant to go north of the Isthmus. Another manifestation was the energetic building in the early 470s of a proper set of walls for the city of Athens, an episode elaborately described by Thucydides to demonstrate the guile of Themistocles, who deceived the Spartans over the affair. Whether the walls were entirely new or a replacement for an Archaic circuit is disputed; Thucydides implies that there was a pre-existing circuit, but no trace of this has been established archaeologically. The Themistoclean circuit, on the other hand, does survive, although the solidity of the socle does not quite bear out Thucydides' dramatic picture of an impromptu "all hands to the pump" operation carried out with unprofessional materials. Sparta's reluctance to see Athens fortified and its anger—concealed but real—after the irreversible event show that even then, despite its cautious attitude to the mainland Ionians, Sparta was not happy to see Athens take over completely its own dominant military role. Or rather, some Spartans were unhappy, for it is a feature of this period that Sparta wobbled between isolationism and imperialism, if that is the right word for a goal pursued with such intermittent energy. This wobbling is best explained in factional terms, the details of which elude the 20th century as they did Thucydides. Thucydides disconcertingly juxtaposes the wall-building episode, with its clear implication of Spartan aggressiveness, with the bland statement that the Spartans were glad to be rid of the Persian war and considered the Athenians up to the job of leadership and well-disposed toward themselves. In fact, there is evidence in other literary sources for the first and more outward-looking policy, such as a report of an internal debate at Sparta about the general question of hegemony, as well as particular acts such as a Spartan attempt to expel Medizers from the Delphic amphictyony—*i.e.*, pack it with its own supporters.

Emerging Athenian independence

One easily identifiable factor in the formation of Spartan policy is a personal one: the ambitions of Pausanias, a young man flushed from his success at Plataea. Pausanias was one of those Spartans who wanted to see the impetus of the Persian Wars maintained; he conquered much of Cyprus (a temporary conquest) and laid siege to Byzantium. But his arrogance angered the other Greeks, "not least," Thucydides says, "the Ionians and the newly liberated populations." These now approached Athens in virtue of kinship, asking it to lead them. This was a crucial moment in 5th-century history; the immediate effect was to force the Spartans to recall Pausanias and put him on trial. He was charged with "Medism," and, though acquitted for the moment, he was replaced by Dorcis. Yet Dorcis and others like him lacked Pausanias' charisma, and Sparta sent out two more commanders. Pausanias went out again to Byzantium "in a private capacity," setting himself up as a tyrant to intrigue with Persia, but he was again recalled and starved to death after having taken sanctuary in the temple of Athena of the Brazen House in Sparta. (The end may not have come until late in the 470s.) The charge was again Medism, and there was some truth to this because the rewards given by Persia to Gongylus of Eretria, one of his collaborators, can be shown to have been historical. There was also a suspicion that Pausanias was organizing a rising of the helots, "and it was true," Thucydides says.

Despite its successes in 479, Sparta, then, was as much a

prisoner of the helot problem as ever, and it could not rely on the loyalty of Arcadia or the Peloponnese generally: Mantinea and Elis had sent their contingents to the Battle of Plataea suspiciously late.

The Delian League

The most important consequence of the successful Greek appeal to Athens was the beginning of the Athenian empire, or Delian League (a modern expression). The appeal to Ionian kinship set the tone for the organization and for much of its subsequent history, though one can fairly complain that this does not emerge strongly enough from Thucydides, who always tends to underreport the religious or sentimental factor in Greek politics.

The Athenians first settled which allies should pay tribute in the form of money and which should provide ships; the details of this assessment were entrusted to the Athenian statesman and general Aristides. Tribute, the need for which was assumed rather than explained, was to be stored at Delos, which would also be the site of league meetings, or synods. Thucydides does not add that the choice of Delos, with its associations with Ionian Apollo, was essentially religious in motivation. Nor does he bring out more than the mercenary or revenge motive of the league (to get redress by devastating the king of Persia's territory). In fact, the mood at the league's founding was positive and solemn, with oaths and ceremonies cementing the act of liberation (478–477). It is unlikely that there was much "small print" to which allies had to subscribe. League meetings were to be held, almost certainly, in a single-chamber organization, in which Athens had only a single vote, though a weighty one; there were perhaps undertakings, subsumed in the general oath taking, about not deserting or refusing military contributions. Unfortunately there are no inscribed stelae, or pillars, as there are for the Second Athenian Confederacy a century later, recording precise pledges by Athens or (equally valuable) listing the members in the order of their enrollment. Apart from the big Ionian islands and some mainlanders, there were in fact Dorian members like Rhodes and Aeolians like Lesbos; there even were some non-Greeks on Cyprus, always a place with a large Semitic component. (Some Cypriot communities probably joined at the outset.) Some Thracian cities were surely enrolled very early. There was no doctrinaire insistence that the league should be exclusively maritime, though the facts of geography gave it this general character automatically. For instance, by mid-century it seems that (in the period of Athens' decade of control of Boeotia, 457–446) the land-locked cities of Orchomenus and Akraiphia were in some sense members. Nor was the league necessarily confined to the Aegean: in 413, financial contributions from Rhegium in the south of Italy, among other places, were handled by the imperial "Treasurers of the Greeks." No inscribed records of tribute exist before 454 BC; after that point, one has the intermittent assistance of the "Athenian Tribute Lists," actually the record of the one-sixtieth fraction paid to the goddess Athena. It should be stressed that until roughly the late 450s there are virtually no imperial inscriptions at all.

Such lack of evidence makes it difficult to show in detail the increasing oppressiveness of the Athenian empire in the second half of its existence (450–404), particularly in the 420s when policy was affected by demagogues like the notorious Cleon. There is simply too little comparative material from the first three decades, and, in the absence of documentary material and of detailed information like that provided by Thucydides for the Peloponnesian War of 431–404, one must infer what happened from the very sparse literary account Thucydides gives for the years 479–439 and from supplementary details provided by later writers. Although it is right to protest, against facile talk of the harsh imperialism of Cleon, that imperialism is never soft, an important but sometimes overlooked chapter of Thucydides is nonetheless explicit that Athens suffered a loss of goodwill through its excessive rigour.

By the middle of the 470s, Greek unity had not come too obviously apart, though the reluctant withdrawal of Sparta was ominous. Even so, at the Olympic Games of 476, an unusually political celebration (the first after the last of the Persian Wars and held in the honoured presence of the Athenian Themistocles), there were still victorious competitors from Sparta, as well as from other Dorian states such as Argos and Aegina and from Italy and Sicily.

Mounting Athenian aggression

Athens' capture of Eion on the Strymon, also in 476, was perfectly in keeping with the ostensibly Panhellenic or anti-Persian program of the Delian League: Eion, an economically and strategically important site in northern Greece, was still held by a Persian commander. This, the first act of the league recorded by Thucydides, was undertaken by Cimon, the son of Miltiades.

The next act of Cimon and the Athenians, the attack on the island of Scyros, was considerably more dubious. Cimon expelled the "Dolopians" (*i.e.*, the indigenous inhabitants) allegedly because they were pirates. Protection against piracy was surely as real a justification for the Delian League as protection against Persia and more general in its application (vulnerability to Persia was very much a matter of geographic position). That Athens was effective in this respect is suggested by the rash of evidence for recrudescent piracy in the early 4th century, when Athens no longer had the power to police the seas. Nonetheless, the treatment of these Dolopians, who were hardly a serious threat to peaceful commerce, certainly appears to have been an act of mere muscle flexing. The enterprise had a propagandist point as well: Cimon brought back the bones of Theseus from the island to Athens, where they were housed in a shrine built for them, somewhere in or near the Agora—perhaps to the east of it. (The site has not been discovered; the so-called "Theseum" is generally agreed to be a temple to Hephaestus.) This magnificent piece of theatre must have been in imitation of the Spartan treatment of the bones of Orestes; this is not surprising, because Cimon was perhaps the first identifiable "Laconizer," or admirer of Spartan values, in Athenian history. Theseus had a special significance not only for Cimon but for the Athenian empire in general. It was Theseus who, according to the myth, had founded the great Ionian festival at Delos called the Delia, which Athens was to revive with much pomp in 426.

More aggression followed, unequivocally directed against other Greeks: Carystus, at the southeastern end of Euboea, was forced to join the league. This was a stepping-up of an Athenian involvement in Euboea that goes back to the 6th century, when Athens installed a cleruchy on Chalcis soon after the Cleisthenic reforms. In the middle of the century inscriptions show that wealthy Athenians possessed land on the Lelantine Plain. Such ownership by individual wealthy Athenians of land in the subject cities of the empire is a telling phenomenon, because the land was usually acquired in defiance of local rules: landowning was normally restricted to nationals of the state in which the land was situated. For Athenians to acquire such land, otherwise than by inheritance as a result of marriage to a non-Athenian, was an abuse, and inheritance of this kind was much less likely after a law of 451 restricting Athenian citizenship to persons of citizen descent on both sides. After 451 "mixed marriages" must have been far less common.

A still more sinister move was the reduction of Naxos, probably in the early 460s. Thucydides equates the inhabitants' loss of freedom with "enslavement"—a strong metaphor. (The precise chronology of the whole period 479–439, and particularly the first 30 years, is uncertain, because Thucydides gives no absolute dates and there are none from other sources before the events in the northern Aegean of 465. The chronology followed here is the orthodox one, but some scholars seek to down-date the attacks on Eion and Scyros to 469—leaving the 470s implausibly empty of known imperial action—and Naxos later still.)

The Battle of the Eurymedon River

The anti-Persian aspect of the league had not, however, been forgotten, in spite of all this activity against Greeks. In 467 Cimon won the great Battle of the Eurymedon River in Pamphylia (southern Anatolia), a naval victory that made a great impression both in Greece (where it was celebrated by the dedication of a bronze date palm, or *phoinix*, at Delphi: a punning reference to the defeated Phoenician fleet) and among waverers, outside Greece proper, who had not yet joined the league. Many new allies were now recruited, such as the trading city of Phaselis on the Lycian-Pamphylian border. A rare early imperial

inscription of the late 460s details the judicial privileges accorded to Phaselis.

Greek success in the east was followed by some mixed achievement under Cimon in the north. A quarrel arose in 465 with the wealthy and fertile northern Aegean island of Thasos about the island's trading stations and mines along the mainland area just opposite it, and Thasos revolted. The word "quarrel" is obviously a euphemism for a piece of naked economic aggression by Athens; all ancient states wished to get their hands on as much precious metal for coinage as possible. Thasos was reduced and forced to give up all of its mines and mainland possessions. A further attempt at this time to extend Athenian northern interest, the colonizing expedition sent to the Nine Ways, the site of the later Amphipolis, was less successful. If silver was one coveted commodity, ship-building timber was another, and the desire for the latter was a large part of Athens' motive for getting a foothold in the Amphipolis region. The Nine Ways operation is a reminder that colonizing activity did not cease with the end of the Archaic period: 10,000 settlers were sent. But the entire force was destroyed at Drabescus. This was probably the occasion for instituting state burial for war dead, a democratic measure that anticipated the reforms at the end of the 460s.

Thasos signaled changes in foreign policy alignments all over Greece. The Thasians had appealed to Sparta for help, asking it to invade Attica, and the Spartans secretly agreed to do so. According to Thucydides, they would have done it had they not been detained by a massive revolt of the helots, who had taken advantage of an earthquake to occupy the strong position of Ithome in Messenia. Ithome, together with the Acrocorinth, the citadel of Corinth, was described by a Hellenistic ruler as one of the "horns of the Peloponnesian ox" that a would-be conqueror had to seize. It is indeed possible that the occupiers of Ithome planned not only an act of secession but, in fact, an attack on the famously unravaged city of Sparta itself. The earthquake not only shook Spartan nerve but must also have had serious demographic effects, though how long-term these were is disputed.

The Spartan response to Thasos looked forward in its anti-Athenian aspect to the great Peloponnesian War of 431–404. It was one of three major episodes in the period up to that war when Sparta moved against Athens. The second was an aborted invasion of Athens under King Pleistoanax in 446. The third episode, in 440, revolved again around the issue of whether to intervene to prevent Athens disciplining a recalcitrant ally, this time Samos. The actual confrontation between Sparta and Athens did not happen in any of these cases. Among the reasons for this—apart from the helot revolt that took a decade for Sparta to put down—was the growing anti-Spartan restlessness in Arcadia. The Athenian Themistocles, who had fallen from favour at Athens and spent time in the Peloponnese after his ostracism (perhaps 471), might have been behind this, though attempts to associate him with particular "synoecizing" developments in the Arcadian cities (*i.e.*, developments whereby small communities coalesced into a single city) are speculative. Nor need such synoecizing (if it happened at this time) necessarily have been democratic and thus evidence that the communities in question were following the Athenian model rather than the Spartan oligarchic one. The evidence of Athenian tragedy (the *Suppliants* of Aeschylus) cannot be pressed to yield secure allusions to Themistocles. Another reason was the continued revival of Argos; its population had now recovered from the defeat at Sepeia (494), and the temporarily exiled descendants of the casualties of Sepeia, the "sons of the slain" as Herodotus calls them, a naturally anti-Spartan group, were now back in control (after ousting the slaves). Argos is on record as fighting a battle in perhaps the 470s, together with Arcadian Tegea, against Sparta, which also had to cope with "all the Arcadians except the Mantineans" at a strictly undatable battle of Dipaieis (which, however, should be put earlier than the Ithome revolt).

The "secret" promise to Thasos was followed by a more open rebuff to Athens. Sparta had invited the Athenians to help with the siege of the helots on Ithome, but with its usual catastrophic indecision Sparta then dismissed the Athenian contingent on suspicion of "revolutionary tendencies." Athens reacted by allying itself with Argos and Thessaly, which was a blow to Spartan ambitions both in its obvious stronghold, the Peloponnese, and in central Greece, an area into which one group of Spartans always seems to have wanted to expand.

This phase of foreign policy has to be somehow associated with internal change at Athens, the so-called Ephialtic reforms. In 462, together with the young Pericles, the Athenian statesman Ephialtes pushed through the decisive phase of the reforms, namely an assault on the powers of the Areopagus. These powers, except for a residual jurisdiction over homicide and some religious offenses, and perhaps a formal "guardianship of the laws," were redistributed among the Council of Five Hundred and the popular law courts. This is, in essence, the very bald and unhelpful account of our main source, the *Constitution of Athens;* there must have been more to it, but the problem is to know how much more. Probably the Areopagus ceased to hear crimes against the state, and such cases were transferred to the popular courts. Alternative interpretations of the inadequate evidence, however, are possible: there are a handful of recorded treason trials earlier than 462 in which a popular element does admittedly play a prominent part, and, although these can be explained away in various ways, it can be held that the transfer of jurisdictional power to the people occurred earlier than 462. Alternatively, it is possible that Ephialtes' reforms in this area involved a mere transfer of "first-instance" jurisdiction (*i.e.*, jurisdiction over cases other than those on appeal) from the Areopagus to the Council of Five Hundred. This requires the assumption of an unattested early 5th-century reform transferring capital appeals to the people.

The reforms of Ephialtes

More radically, and generally, the jurisdiction of magistrates (archons) was much curtailed; they now conducted a mere preliminary hearing, and the main case went to a large popular jury. The authority to conduct inquiries into the qualifications for office of the archons themselves (the *dokimasia* procedure) and into their behaviour after their terms of office had expired (*euthyna* procedure) was also taken away from the Areopagus and given to the Council of Five Hundred. This principle of popular accountability seems new, though the statement in Aristotle's *Politics,* that the right of popular *euthyna* goes back in some sense to Solon, has its defenders.

There surely were other reforms. Certain features of the later democracy appeared after the rule of Cleisthenes but were in place by the Peloponnesian War; it is plausible to argue that they were introduced at this time, though there is a risk of circularity in characterizing Ephialtes as a comprehensive reformer by reference to strictly unattributed and undated changes. Thus, sortition for the Council of Five Hundred is not likely to have been earlier than 487, when the archonship ceased to be elective; but Athens imposed sortition for a comparable though smaller council on Ionian Erythrae in 453, surely not before there was sortition for the Council at Athens itself. Similarly, there is evidence for jury pay for the 460s (or less probably for the 450s), which makes it plausible to date Council pay, attested by 411, to the mid-century period also.

Taken together these reforms look like the result of careful thinking by particular individuals with a definite democratic philosophy. A case, however, can be made for seeing them all as part of a 30-year process, with a central action-filled phase, rather than as a single event. After all, the Areopagus was affected indirectly by the changes in the archonship in 487, though the archonship was formally opened to the *zeugitai* (the hoplite class) only in 457. But despite the great increase in work for the big popular juries and the granting to the courts of the right (which may go back to Ephialtes) to quash or uphold allegedly unconstitutional proposals, it is not likely that then or at any other time Athenians saw themselves as conferring sovereignty on the people's courts at the expense of the Assembly. The implied psychological distinction between juries of Athenians and political gatherings of the same Athenians is not a plausible one.

Some of these changes were perhaps already in the air when the Spartans dismissed Cimon and his Athenians at Ithome. Cimon's absence seems to have given Ephialtes and Pericles their chance: the main Areopagus reform was passed at this time, and in 461 Cimon was ostracized. This rejection of Cimon, however, was a personal matter: he should not be seen as a "conservative" opponent of a reform that gave more power to the people and especially to the thetic class, which manned the fleet. For one thing, Cimon's victory at the Battle of the Eurymedon River was primarily a naval victory; for another, it was the Sparta-loving Cimon and his hoplites who were dismissed by the Spartans from Ithome for their subversive tendencies. Most important of all, there is the general point that the interests of hoplites and *thētes,* now as at other normal times, coincided; both were denied the chance of standing for the archonship before 457 (the hoplites were admitted to it in that year). On the whole, it is the top two "Solonian" groups, the *pentakosiomedimnoi* and the cavalry class who were bracketed together on the one hand (as by Thucydides in one military context), while the *zeugitai* and *thētes* tended to be bracketed together on the other. No built-in class cleavage existed between the hoplite or zeugite class and the *thētes,* and attempts to exploit one, by advocating or offering a "hoplite franchise," were short-lived failures. Cimon then should not be seen as champion of "conservative" hoplites against "radical" *thētes;* this view is wrong because the interests of hoplites and *thētes* were indissolubly linked.

Athens' two new alliances, with Argos and Thessaly, were provocative (surely not just defensive), but they did not create direct danger of war. Far more serious was the friction at this time between Athens and Corinth. Corinth had made no move to help Sparta, as far as is known, at the time of the Ithome disaster but seems to have pursued expansionist goals of its own in the Peloponnese, perhaps at Argos' expense. Now that Athens and Argos were allied, this indirectly tended to damage Corinth's hitherto good relations with Athens. (Corinth had fought well at Salamis, as even Herodotus was aware, though very different stories were circulating on this topic after 460.) More relevant than this was Athens' ready reception of a third ally, Megara; like the Argives, the Megarians had also felt pressure from Corinth (one hears of a boundary dispute and a local war) and turned to Athens. This was the cause and beginning of the "violent hatred" between Corinth and Athens, which produced what modern scholars call the First Peloponnesian War.

Causes of the First Peloponnesian War

The First Peloponnesian War (460–446) should probably be seen as essentially a conflict between Athens and Corinth, with occasional interventions by Sparta. Modern disagreement centres on the reasons why Sparta did not play a role: one line of explanation is purely military, invoking the difficulty of invading Attica while the mountains above Megara were policed by Athens; the other and more plausible view is that Sparta simply lacked the will to act consistently. Spartan inactivity should in any case not be exaggerated. There is a pattern to its interventions, which suggests that in this period, as at others, the "central Greek" lobby at Sparta, the closest thing to an identifiably imperialistic group to be found there, could sometimes prevail.

The first battle of the war, at Halieis in the Gulf of Argolis, was a Corinthian victory, but the next battle, at Cecryphalea (modern Angistrion), went Athens' way (459). Aegina, which was attacked and besieged in the same year, was reduced in the following year and forced to pay tribute, though some vague undertaking about autonomy may have been made; the alternative is to suppose a special clause about Aeginetan autonomy, or even a general autonomy clause, in the peace of 446, which ended the war. The alleged Athenian infringement of the autonomy of Aegina was one of the secondary causes of the main Peloponnesian War. In the meantime, the subjugation of Aegina, a great city of the Archaic age, whose proud Dorianism and traditions of seafaring and hospitality are stressed in lines of great beauty by Pindar in his *Nemean Odes* and elsewhere, was an event of cardinal importance. The pretense that Athens was merely leading a voluntary association of willing Ionian cities in need of protection could hardly survive the reduction of Aegina.

The real scale of Athenian ambitions is shown by four other developments of this period. First, Athens undertook a great and disastrous expedition to Egypt (460–454), in ostensible continuance of the fight against Persia. Egypt, however, had always been a rich and desirable Persian satrapy, exploited by absentee Persian landowners; and thus an economic motive for Athens cannot be excluded. Second, Athens made an alliance (almost certainly in 457) with an inland half-Greek Sicilian city, Segesta. This prepared the ground for a more tangible western policy in the 440s. Third, Athens now built the Long Walls connecting it to Piraeus and thus the sea and making it possible to depend for the future on the produce of its empire if absolutely necessary. The walls, however, should not be thought of as purely defensive in view of the constant connection made by Thucydides between walls and dynamic sea power. Fourth, Athens made an alliance (the inscription is strictly undatable) with the Delphic Amphictyony in the middle of the century. This must be connected with the Athenian alliance made with Thessaly (see above) in 461, because Thessaly controlled a majority of Amphictyonic votes (always a reason why other states or rulers, like Philip of Macedon in the next century, were anxious to have a controlling interest in Thessaly). It is interesting that Athens should thus extend its religious propaganda to include the sanctuary of Apollo of Delphi (Apollo Pythios) as well as that of Apollo of Delos. The oracle was always a distinct entity from the sanctuary, but it cannot be accidental that about now the oracle, normally favourable to Sparta in this period and conspicuously so in 431, declared Athens an "eagle in the clouds for all time."

Athenian expansion

The central Greek line of Athenian expansion was bound to bring a collision with Sparta. It entered the war in 458 in response to an appeal by its "mother city" Doris, the city from which the primeval Dorians were believed to have set out to undertake the invasion of the Peloponnese. This tiny state in central Greece was currently experiencing difficulties with its neighbour Phocis. The religious and sentimental factor in Sparta's response was not negligible, but Sparta may have had other aims as well. There is evidence in Diodorus Siculus, a Greek historian of the 1st century BC, though not in Thucydides, that Boeotia was a target.

It was on their return from Doris that the Spartans finally came to blows with Athens at Tanagra in Boeotia (458). The battle was of large scale—one hears of Argive involvement on the Athenian side—but indecisive. The Athenians, however, followed it up with a victory at Oenophyta, which gave them control of Boeotia for a decade, an extremely important development passed over by Thucydides in half a dozen words. There was further aggressive Athenian action, first under the general Tolmides, who circumnavigated the Peloponnese (456) and perhaps settled the large number of Messenians at Naupactus alongside the original Naupactans, and second, under Pericles, who launched military expeditions in the Gulf of Corinth (454?). But the disastrous end to the adventure in Egypt (454) made Athens ready for a truce, and in 451 Athens came to terms with Sparta, while Argos concluded a 30-years' peace with Sparta on its own account.

Athens resumed the war against Persia with hostilities on Cyprus, but Cimon's death there made diplomacy imperative in this sphere also. This is where one should place the Peace of Callias (449), mentioned by Diodorus but one of Thucydides' most famous omissions. Thucydides' subsequent narrative of the Peloponnesian War, however, presupposes it at a number of points, especially in the context of Greek dealings with Persia in 411. More generally, a peace is made likely by the history of the 440s and 430s, which records no more overt Athenian warfare against Persia and a certain restlessness inside the Athenian empire. (This absence of warfare may be due to other factors as well; it is possible that the Treasury of the League, to which various states in the Delian League paid tribute, was moved from Delos to Athens in 454, a centralizing gesture that may have caused alarm. But the move may have happened earlier.)

Peace of Callias

Nonliterary evidence also points in the direction of a peace: the evidence of inscriptions makes it probable that no tribute was levied in 448. Perhaps it was recognized that the struggle with Persia was over and with it the justification for tribute; if so, the recognition was only momentary, because there was tribute again in 447. Furthermore, an inscription of the 420s appears to refer to a renewal of the peace on the death of Artaxerxes I. Finally, the commissioning of a new Temple of Athena Nike ("Victory"), and perhaps even of the Parthenon, may have been an aspect of the same mood. (The peace could be represented as a victory of a sort because it restricted the Persian king's naval movements.) Yet the close correlation of architectural with political history is to be avoided; antibarbarian artistic themes on Greek public buildings need no special explanation at any time in the 5th century. Against all this there are a few ancient allegations that the peace was a later forgery, an implausible idea because such diplomacy was a matter of public knowledge.

Despite the truce with Athens in 451, Sparta had not withdrawn into its Peloponnesian shell completely. In addition to its campaign in support of Doris, Sparta successfully intervened in central Greece in a "Second Sacred War" against Phocis, which, with the assistance of Athens, had gained control of Delphi. Sparta handed over Delphi to the Delphians, but this action was promptly neutralized by the Athenians, who restored the sanctuary to Phocis. Catastrophic revolts in Boeotia and Euboea (446), however, soon eroded that Athenian authority in central Greece of which the Delphic intervention was a manifestation.

The tributary states had much cause to rebel. There was something ominous about the sheer physical scale of the first (in chronological order) of the stone blocks on which were carved, as a permanent record, the tribute payments due to Athena. The block, preserved in the Epigraphic Museum in Athens, is a towering 142 inches (3.61 metres) high and had plenty of room for many years of tribute. Evidently the Athenians of 454 expected the empire to go on indefinitely, despite the failure in Egypt, which must have made many observers reflect that peace with Persia could not be far away. Yet tribute, exactingly collected, as Thucydides says, was not the only grievance. It was not even the only economic grievance. In the period of the early Peloponnesian War there were, as inscriptions show, strict Athenian controls on the traffic of grain from the Black Sea, including "Guardians of the Hellespont." According to one view, these controls were a purely wartime expedient, but, given the state of the evidence, that charitable view is an abuse of the argument from silence; in any case, a prewar inscription does in fact attest a 10 percent tax on shipping from the Black Sea. Grain bound for Athens itself was probably exempt from this. Still in the economic sphere, resentment against Athenian ownership of land—whether collectively (the so-called cleruchy system, stepped up at the end of the 450s) or privately, by wealthy individuals—can legitimately be inferred from the self-denying promises made by Athens in the days of its 4th-century confederacy. In this category should be included sacred precincts (*temenē*) in allied states, marked out by *horoi*, or boundary stones, which indicated land that might be leased out to other wealthy Athenians.

Another interference in the internal affairs of tribute-paying allies in the 4th century was the placement of garrisons and garrison commanders, attested as early as the Erythrae decree of 453. The same decree imposed a "democratic" constitution, according to a principle that the literary sources say was general Athenian policy. Yet it would be simplistic to think that such Athenian-influenced constitutions were necessarily a significant upholding of human rights. One must always ask what "democracy" can have meant in a small community. At Erythrae, not only was the council less democratic than that at Athens (see above), but there also was a property qualification for jurors. And at exceedingly few places other than Athens does inscriptional evidence for amendments from the floor exist. In any case, there are significant exceptions (Samos, Mytilene, Chios, Miletus, Potidaea, and possibly Boeotia) to the generalization that Athens insisted on democracies.

The Erythrae decree

What the allies thought of this is inscrutable. A statement by an Athenian speaker in Thucydides that the popular party everywhere supported Athens is matched by the reported view of another Thucydidean speaker that what the allies wanted was freedom from interference of any kind.

In the legal sphere the allies suffered from disabilities (such as the requirement to have certain types of cases heard in Athens). These were firmly maintained even in texts, such as the Phaselis decree, that accord specific limited legal privileges. Full legal privilege and status was reserved for full Athenians, a status whose definition was tightened by the citizenship law of Pericles in 451. Roman commentators pointed to Athenian (and Spartan) failure to integrate their subjects as citizens as the explanation of their more general failures as imperial powers. There is much in this; it is not an answer to say that there is no attested clamour for Athenian citizenship, when the allied view on so many points does not exist. Certainly, among the thousands disfranchised in the 440s by the new rules regarding citizenship, there must have been many immigrants from the empire. Colonial mother cities sometimes offered citizenship wholesale to their daughter communities. Imperial Athens borrowed many features of the colonial relationship, but not that one.

Boeotia revolted in 446 with help from Euboean exiles, and Athens was forced to accept this political reversal after a military defeat at Boeotian Coronea. The revolt of Euboea itself followed. Pericles crossed over to deal with it but only precipitated a third revolt, that of Megara. This was a serious military crisis, and it was compounded by a Spartan invasion of Attica: King Pleistoanax got as far as Eleusis and the Thriasian plain, but, as mentioned above, the invasion was not carried through. Pleistoanax and Pericles seem to have struck a deal: Sparta would not interfere in Euboea or invade Attica in exchange for Athens' acquiescence in the loss of Boeotia, the Megarid, and certain Peloponnesian sites. An arrangement on these lines was formalized in a Thirty Years' Peace between Athens and Sparta, but it would be too optimistic to try to list all the terms. An essential undertaking was a renunciation of armed attacks if the other side was prepared to submit to arbitration. Athens could now deal with Euboea, and inscriptions have preserved the terms of the firm settlement imposed on individual communities there.

Athenian buoyancy was not deflated even by these failures. For in 443 it advertised a big colonial venture to Thurii in Italy and about the same time made alliances

From J. Boardman, J. Griffin and O. Murray (eds.), *The Oxford History of the Classical World*, © Oxford University Press, 1986



The Athenian Empire at its greatest extent.

with Rhegium in Italy and Leontini in Sicily (alliances renewed a decade later on surviving inscriptions).

Western Greek communities

Since the Persian Wars the most splendid of the western Greek communities had been tyrannically ruled until the fall of Gelon's family, the Deinomenids of Syracuse, in 466/467, soon after the death of Gelon's brother Hieron and the fall of the tyrannical house of Theron at Acragas in 472. Syracuse enjoyed a moderate democracy thereafter, disturbed only by a native rebel, Ducetius, whom it took surprisingly long to put down. In Italy, where Rome was preoccupied with the neighbouring Volsci and Aequi for much of the century, Hellenism maintained itself vigorously: the temples of Paestum dating to the 5th century, like those of Acragas or Segesta, were comparable to anything in mainland Greece, and there were philosophers and the philosophical schools of Croton, Taras, and Elea (Velia), all in southern Italy. At Elea, south of Paestum, interesting portrait statues were discovered in the 1960s, which showed that the philosophical school there had a medical aspect to it: a cult of Apollo Oulios, a healing god, was looked after by a clan of Ouliadai (which was associated with the medical organization, though the exact relationship is obscure), and even the famous Parmenides, better known as a philosopher, is called Oulíades. Pythagoreanism, a philosophical school and religious brotherhood, flourished in southern Italy. In the early 5th century Pythagorean groups involved themselves in government, ruling Croton for a period. Nonetheless, there were tyrannies in southern Italy too, such as that of Anaxilas at Rhegium. Religious and social links with the Greek mainland were cultivated, above all by contacts with the sanctuary and games at Olympia and by patronage of poets like Pindar.

The Athenian-inspired Thurii project represented a fairly substantial mainland Greek encroachment on western soil; this and a mysterious Athenian colonizing effort in the Bay of Naples region, undertaken perhaps in the early 430s by a western expert, Diotimos, must have caused unease to western-oriented Corinth. (There is even a Spartan aspect: Thurii was soon engaged in warfare with Sparta's only historical colony, Taras.) Nonetheless, when Samos revolted from Athens in 440, it was Corinth that in a congress of the Peloponnesian League voted against intervention against Athens on behalf of Samos (Corinth's attitude had no doubt softened with the detachment of Megara from Athens). Sparta, however, seems to have wanted to stop Athens in its tracks, though in the end it was typically unwilling to press this line of policy home. At this point Thucydides' main narrative stops for five crucial years, at the end of which tension between Corinth and Athens was again high, on the eve of the great Peloponnesian War.

The Peloponnesian War. The causes of the main Peloponnesian War need to be traced at least to the early 430s, although if Thucydides was right in his general explanation for the war, namely Spartan fear of Athenian expansion, the development of the entire 5th century and indeed part of the 6th were relevant. In the early 430s Pericles led an expedition to the Black Sea, and about the same time Athens made an alliance with a place close to areas of traditional Corinthian influence, Acarnania. (On another view this belongs in the 450s.) In 437 Athens fulfilled an old ambition by founding a colony at Amphipolis, no doubt on a large scale, though figures for settlers do not exist. This was disconcertingly close to another outpost of Corinthian influence at Potidaea in the Chalcidice, and there is a possibility that Athens subjected Potidaea itself to financial pressure by the mid-430s. That city was an anomaly in being both tributary to Athens and simultaneously subject to direct rule by magistrates sent out annually by Corinth; it clearly was a sensitive spot in international relations. Thus to the west (Acarnania and other places) and northeast (Amphipolis, Potidaea) Corinth was being indirectly pressured by Athens, and this pressure was also felt in Corinth's own backyard, at Megara. Athens passed a series of measures (the "Megarian decrees") imposing an economic embargo on Megara for violations of sacred land. The religious aspect of the offense was reflected in the exclusions imposed: like murderers, the Megarians were banned from the Athenian marketplace and the harbours in the Athenian empire. But one should not doubt that Athens caused and intended to cause economic hardship as well or that the decrees were the first move in securing Megara as a military asset, a line of policy further pursued in the years 431 to 424.

The Megarian decrees

Reactions to all this, within the empire and outside it, are hard to gauge. Athens' savage reduction of Samos, a member of the Delian League, in 440–439, did not stop Mytilene and most of Lesbos from appealing at some time in the prewar period to Sparta for encouragement in a revolt they were meditating. No encouragement was given: Sparta was standing by the Thirty Years' Peace and should be given (a little) credit for doing so.

For the period from 443 to 411 a vastly more detailed narrative is possible than theretofore, but the reader should be warned that this freak of scale is due to one man, Thucydides, who imposed his view of events on posterity. It would, however, be artificial to write as if the information for this unique period were no better than that available for any other.

The main precipitating causes of the war, thought of as a war between Athens and Sparta, actually concerned relations between Sparta's allies (rather than Sparta itself) and other smaller states with Athenian connections. The two "causes" that occupy the relevant parts of Thucydides' first, introductory book concern Corcyra and Potidaea. (Thucydides does not let his readers entirely lose sight of two other causes much discussed at the time—the Megarian decrees and the complaints of Aegina about its loss of autonomy. One 4th-century Athenian orator actually dropped a casual remark to the effect that "we went to war in 431 about Aegina.") Corcyra, which had quarreled with Corinth over the Corcyran colony of Epidamnus on the coast of Illyria (a colony in which Corinth also had an interest), appealed to Athens. Taking very seriously the western dimension to its foreign policy (it was about then that the alliances with Rhegium and Leontini were renewed), Athens voted at first for a purely defensive alliance and after a debate, fully recorded by Thucydides, sent a small peace-keeping force of 10 ships. This was, however, trebled, as a nervous afterthought; no political background is given for this move, which, moreover, emerges only subsequently and in passing during the narrative of events concerning Corcyra itself. (This is a small illustration of the important point that Thucydides' presentation unduly influenced modern views on the general issue of Athenian belligerence, as on so many other issues. A different narrative, by emphasizing the escalation of the Athenian commitment and making it the subject of another full debate, might have left a different impression.) In fact, Corinthian and Athenian ships had already come to blows before the reinforcements arrived.

Then at Potidaea, a Corinthian colony, the Athenians demanded that the Corinthian magistrates be sent home. Potidaea revolted, and an unofficial Corinthian force went out to help. Potidaea was laid under siege by Athens. None of this yet amounted to war with the Peloponnesian League as a whole, but the temperature was as high as it could be, short of that. A congress of Spartan allies was convoked to discuss grievances against Athens, and the decision was taken for war.

The other Spartan ally seeking to involve Sparta in a private feud with an enemy was Thebes, whose attack on its neighbour Plataea (an Athenian ally) in time of peace was retrospectively recognized by Sparta as an act of war guilt. Sparta should not have condoned it, nor should it have invaded Attica (despite the fact that Athens had placed a garrison in Plataea) so long as Athens was offering arbitration, as it seems it was. Thucydides vacillates between two events for the beginning of the war, the invasion of Plataea and the Spartan invasion of Attica. Both occurred in 431, separated by a mere 80 days.

Pericles

Athenian war strategy and the initial conduct of the war are presented by Thucydides very much in personal terms: the focus is on what Pericles, the dominant figure of this time, did or wanted. This method, like the Homeric emphasis on heroes, is to some extent literary spotlighting, for at no time was Pericles immune from criticism. In the 440s he had to deal with a major rival, Thucydides, son

of Melesias (not the historian), who was ostracized in 443. Even after that, in the poorly documented 430s (before Aristophanes and Thucydides provide information about individual figures of second- or third-rate significance), there are suggestions of tension, such as a partial ban on comedy (with its potential for exposure) and indications in the sources that Cleon was really not a successor of Pericles at all but a highly critical contemporary. The reasons for Pericles' ascendancy remain a secret, and this in itself makes it necessary to allow for a large element of "charismatic" leadership.

In the military sphere Thucydides is surely wrong to present Pericles as a one-man band. He says of Pericles that early in the war "the Athenians reproached him for not leading them out as their general should." If this sentence had survived in isolation, one would hardly have guessed that Pericles was one of the college of 10, subject to control and threat of deposition by the Assembly (Pericles was indeed deposed temporarily toward the end of his life). On the whole, however, Thucydides minimizes the degree to which Athenian generals enjoyed executive latitude, particularly in wartime; it may be suggested that the reason for this was his own exile, imposed in 424 as a punishment for failing, as commander in the region, to relieve Amphipolis. This impressed him deeply—and unduly—with the impotence and vulnerability of generals other than Pericles.

The reproach of "not leading out the Athenians" provides useful insight into Periclean strategy, revealing it to have been largely reactive. Whereas Sparta's goal was to liberate Greece from tyranny, which required it to dismantle the Athenian empire, all Athens had to do was to avoid such demolition. In a way this suited neither side: initiative of the kind this demanded from Sparta was in short supply there (though never entirely absent). For the Athenians' part, the famously energetic and meddlesome population did not take kindly to the practical consequences of Periclean strategy that required it to evacuate Attica and move its population behind the fortified walls of Athens, to rely on accumulated capital reserves and on the fleet as an instrument to hold the empire firmly down, and to avoid adding to the empire during wartime. By these means the Athenians would eventually "win through" (the Greek word is neatly ambiguous as between victory and survival).

Actually the Athenian position was not and could not be so simple. For one, the agricultural evacuation of Attica was not as complete as it was to be after 413 when the Spartans occupied Decelea in northern Attica. Nor did Pericles altogether abandon Attica militarily: there were cavalry raids to harass the dispersed foot soldiers of the enemy and to keep up city morale. Holding the empire down and holding onto capital were potentially inconsistent aims in view of the great cost of siege warfare (there was no artillery before the 4th century to facilitate the taking of fortified cities by storm). The destruction of Samos had been expensive—a four-figure sum in talents—and the siege of Potidaea was to cost 2,000. Athens, even with coined reserves of 6,000 talents at the beginning of the war, could not afford many Potidaeas. Pericles can be criticized for not foreseeing this, with the evidence of Samos behind him.

The cost of siege warfare

Sparta came as a liberator. This too called for money and ships, but it had neither accumulated reserves like Athens nor a proper fleet. Persia was a possible source for both, but assistance from Persia might compromise Spartan "liberation theology." This was especially true if Sparta set foot in Anatolia, where there were Greeks with as much desire for liberation (whether from Athens or Persia or both: some communities paid tribute in both directions) as their mainland counterparts. A further difficulty lay in the kind of regime Sparta itself could be expected to impose if successful. One revealing reason for the failure of the big colony at Heraclea founded in 426 (see below), a project with a strongly anti-Ionian and propagandist element, was the harsh and positively unjust behaviour of the Spartan governors, who frightened people away.

Again a few qualifications are in order. Money could be obtained from more acceptable sources than Persia—from the western Dorians, for instance. And subsidized piracy, of which one hears a little in the 420s, was another solution to the naval problem. Against harsh governors like those at Heraclea one has to balance Brasidas, who was as good a fighter in the battle for the hearts and minds as in the conventional sense.

Sparta's invasion of Attica set the tone of the first half of the Archidamian War (431–421), named after the Spartan king Archidamus II, unfairly in view of the wariness he is said to have expressed at the outset. Athens moved its flocks from Attica across to Euboea, whose economic importance was thus raised further still. As if in recognition that this was a war brought about at the instance of Corinth, much early Athenian naval activity was devoted to stripping Corinth of assets in the northwest—of Sollium, Astacus, and Cephellenia. Yet there was also an Athenian raid on Methone in Messenia (the later Venetian strong point of Modon), foiled by Brasidas; a morale-boosting raid on the Megarid (such raids were repeated twice a year until 424); and some successful diplomacy in the north, where the Odrysian Thracians were won over. At the end of this first campaigning year, Pericles delivered an austere but moving speech honouring the fallen men, which has become known as the funeral oration of Pericles. This famous oration, however, is largely the work of Thucydides himself; it is a timeless personal tribute to Athenian power and institutional strength but not, as has been argued, a key to unlock Athenian civic ideology. The speech, as preserved, is not peculiarly enthusiastic about democracy as such and has perhaps been over-interpreted in the light of Athens' later cultural fame. In particular, the Thucydidean Pericles is usually taken to have said that Athens was an education to Greece, but in context he says merely that other Greeks would do well to profit from its political example.

The second year of the war, 430, began with another invasion of Attica. Thucydides, having scarcely brought the Peloponnesians into Attica, switches styles dramatically to record the outbreak of a dreadful plague at Athens. Although it cannot be securely identified with any known disease, this plague carried off one-third of the 14,000 hoplites and cavalry (there was a recurrence in 427). Pericles himself came down with the disease and died in 429, not, however, before leading a ravaging expedition against Epidaurus and other Peloponnesian places and defending himself against his critics. The speech Thucydides gives him for this occasion is as fine as the funeral speech, which has received so much more attention. It hints loftily at expansion to east and west of the kind that Pericles' initial strategy had appeared to rule out. It is possible that this speech is historical and that the purpose of attacking Epidaurus was to bar Corinth's eastern sea-lanes completely; Aegina had already been evacuated and repopulated by cleruchs in 430, perhaps as an initial step toward this end. In the north, Potidaea surrendered, and a cleruchy was installed here too, a further Corinthian setback.

Peloponnesian pressure on Plataea was stepped up in 429. A large expedition in the northwest under the Spartan Cnemus, who used barbarian as well as Greek forces in an effort to win back some of Corinth's losses, showed that there were adventurous thinkers before the northern operations of Brasidas later in the decade. It was, however, a failure, as was a Peloponnesian embassy to Persia asking for money and alliance. Intercepted by the king of the Odrysians, the ambassadors were handed over to Athens, where they were put to death with no pretense at trial. The Odrysians feature prominently at this time (but perhaps Thucydides' own family interests in Thrace have distorted the picture): the mass mobilization of a large Odrysian force, ostensibly in the Athenian interest, soon afterward caused general terror in Greece, but it came to nothing. There was more concrete encouragement for Athens in some naval successes of the great commander Phormion in the Gulf of Corinth.

It is perhaps surprising that it was only in 428 that a revolt within the Athenian empire gave Sparta the opportunity to implement its basic war aim of liberating Greece. This was the revolt of Mytilene on the island of Lesbos, to which Athens reacted with a prompt blockade. It was a shrewd Spartan move to summon the Mytileneans and

The revolt of Mytilene

other injured Greeks to the Olympic Games at this point, thus emphasizing that one aspect of the war was the tension between Dorians and Ionians. (Athens was hardly formally excluded from the solemnities, but Olympia always had a Dorian flavour.) Alcidas, the Spartan commander sent to assist the Mytileneans, failed, however, to do anything for them. On its surrender (427) the city narrowly escaped the wholesale executions and enslavements Cleon had recommended, but only as a result of second thoughts on the part of the Assembly (these events and decisions form the context of the famous "Mytilene debate"). It is to the Athenians' credit that some of them were moved by the thought that their original decision was bloodthirsty.

There were no such doubters among the Spartans who supervised the final phase of Plataean 5th-century history. When the remaining Plataeans surrendered (some had already broken out to Athens), they were put to death to a man, after the "brief question" had been put to them, "Have you done anything for Sparta during the war?" This was a question that the Plataeans, despite some moving pleas, could answer only negatively. At least Cleomenes I in the 6th century and Agesilaus II in the 4th, both of whom applied much the same criterion as this in international affairs, made no pretense of being liberators of Greece. It is impossible for the modern reader to reflect on these two fully reported incidents at Mytilene and Plataea without coming to some general conclusions about Spartan behaviour; and Thucydides, too, was prompted to generalize in this fashion. His thoughts are attached to an account of civil strife at Corcyra, in the west, in 427. After a bloodbath, the democratic pro-Athenian faction prevailed over the oligarchical pro-Spartan party, with the Athenian commander Eurymedon making no attempt to stop it.

About this time the Athenians speculatively pursued their western interests, sending at first an expedition of 20 ships under Laches and Charoeades (*c.* 427) and then 40 more under Sophocles (not the tragedian), Pythodorus, and Eurymedon (426–425). This was a large force in total, given Athens' other commitments, but its goals are difficult to assess; both radical and conservative motives are given, such as the desire to give the sailors practice (not a ridiculous motive, but an inadequate one), to cut off grain shipments to the Peloponnese (by which Corinth is presumably meant), or even to see if the whole island of Sicily could be brought under control, whatever exactly that might entail. (In 424, after mostly halfhearted warfare, the Sicilians put aside their internal differences at a conference in Gela, of which the Pan-Sicilian Hermocrates was the hero. The Athenian commanders returned home to an undeserved disgrace: their mandate for outright conquest had hardly been clear, nor were their resources sufficient.) The attempt by the Athenian general Nicias to take Megara by military means (427) had more immediate promise of success.

It is possible that even the Spartans were uneasy at what the main events of 427, at Mytilene and Plataea, had done for their image: they had been ineffective and brutal. Perhaps in partial redress, but also in pursuit of a traditional line of policy, they issued a general invitation to participate in a large (10,000 strong) colony at Heraclea in Trachis at the southern approach to Thessaly. This colonizing effort had intelligible short-term military motives, namely, a felt need to gain a hold on the Thracian region—the only part of the Athenian empire reachable by land—and a desire to deny Athens access to its larder on Euboea. But Thessaly had always featured and was always to feature in ambitious Spartan thinking; and Sparta may already have had designs on the amphictyonic vote that one certainly finds Heraclea exercising in the 4th century. From the propaganda point of view, the exclusion of Ionians, Achaeans, and some others was telling. Sparta was presenting itself as a leader of Dorians, not just as a selfish promoter of Spartan interests. This was the redress offered to a Greek world well-disposed toward Sparta at the beginning of the war but now perhaps dismayed by the way things were going. It was a pity that the brutality of Spartan governors at Heraclea helped to ruin the project.

Athens' magnificent refounding, also in 426, of the Ionian festival of Apollo on the island of Delos, where the Delian League had been established in 478 BC (see above *The Athenian empire*), must surely in part be seen as a response to Dorian Heraclea. (There were other motives too, such as desire for expiation for the plague, which had ravaged Athens a second time in the winter of 427–426.) Of the two great Panhellenic sanctuaries, Olympia had taken an ugly anti-Athenian look in 428, while the oracle of Delphi had actually approved the Heraclea colony. Athens, through Delos, was creating or inflating religious propaganda possibilities of its own. The same is true of an Athenian invitation to the Greeks at large, also (possibly) in the 420s, to bring offerings of firstfruits to Eleusis.

Land operations in the northwest occupied much of the purely military history of 426. They were conducted by one of the finest generals of the Peloponnesian War, the Athenian Demosthenes (no relation of Philip's 4th-century opponent). He was at first spectacularly unsuccessful in some ambitious campaigning, perhaps not sanctioned by the Assembly at all, in Aetolia, where his hoplites were nearly helpless against the light-armed tactics of the locals. He was, however, able to retrieve the position subsequently, in Amphilochia, in circumstances that brought further discredit on Sparta, whose commander deserted his Ambracian allies. Demosthenes

The decisive year in the Archidamian War was 425. Demosthenes, whose credit with the Assembly must now have been excellent, obtained permission to use a fleet round the Peloponnese. He and his troops used it to occupy the remote Messenian headland of Pylos, a prominence at the north end of the Bay of Navarino, and to fortify it. The Spartans foolishly reacted by landing a hoplite force on Sphacteria, the long island to the south of Pylos. This force of 420 men, about half of them full Spartan citizens, was cut off by the Athenians, who thus acquired a potentially valuable bargaining chip. Sparta sued for peace without reference to its allies (so much for liberation), but Cleon persuaded Athens to turn the offer down. Cleon made steep demands, including (in effect) the cession of Megara, showing that he—like Nicias in 427 and Demosthenes and Hippocrates in 424—grasped the strategic importance of Megara, even if the historian Thucydides did not.

Thucydides disliked Cleon, as did another highly articulate contemporary, the playwright Aristophanes (see in particular his comedy *Knights,* of 425–424). The picture that emerges from their works of Cleon and figures like him as "new politicians," arising not from among the old or property-holding families but from the people, is largely a literary fiction. It was foisted on posterity by these ancient writers, who exaggerated the contrast between Pericles and his successors because they admired Pericles' style. In social background, political methods, and particular policies the difference was not great. The real change in Athenian politics came only with the loss of the empire in 404 and the resulting partial breakdown in the "consensus politics" that had prevailed hitherto (because all social classes stood equally to gain from the empire, which financed political pay, provided land for all, and cushioned the rich against the cost of furnishing the fleet).

There are two lines of policy one can safely associate with Cleon from evidence other than that of Thucydides. One is a large increase in the level of allied tribute (425–424) documented by an inscription. Inscriptions also show the necessity of the raise, proving, as they do, a serious shortfall in public finance (extensive "borrowing from Athena"—*i.e.,* drawing on capital reserves). The other line of policy is an attempt, attested by Aristophanes, to draw Argos into the war in some way (its peace with Sparta was due to expire in 421, the year in which, unknown to Cleon in 425, the Archidamian War was to end).

By declining the diplomatic solution, Cleon found himself committed to a military one. He succeeded dramatically, capturing 120 full Spartans and taking them back to Athens. This operation, achieved partly with the use of light-armed troops, ensured that there would be no invasion of Attica in 424. Athens was free to establish a base on the island of Cythera south of Laconia and make a serious and initially successful attempt on Megara.

At this point the balance of the war began to tilt again

Brasidas' gains for Sparta

in Sparta's favour: Brasidas arrived, on his way to the north, and saved Megara by a whisker. Moreover, an ultra-ambitious Athenian attempt to reinstate the mid-century position by annexing Boeotia failed at Delium; this was a major defeat of Athens by a Boeotian army whose key component was Theban. Meanwhile, Brasidas had reached the north, where he had won over Acanthus by a blend of cajolement and threats and where, too quick for Thucydides (the historian) to stop him, he had taken Amphipolis. From there he proceeded to capture Torone. An armistice between Athens and Sparta in 423 did not stop further northern places from falling into Brasidas' arms—almost literally: at Scione the inhabitants came out to greet him with garlands and generally received him "as though he had been an athlete" (a rare Thucydidean glimpse of a world other than war and politics). He briefly won over Mende as well, but Athens recovered it soon after; Cleon arrived in 422 and won back Torone too. The deaths of both Cleon and Brasidas in a battle for possession of Amphipolis removed two main obstacles to the peace that most Spartans had been wanting for several years—in fact, since Sphacteria. The imminent expiry of the Argive peace was another factor, as was the occupation of Cythera, which provided a base for deserting helots (it is surprising that Athens did not make more use of the Spartan fear of their helots, a far from secret weapon of war). The essence of the Peace of Nicias (421) was a return to the prewar situation: most wartime gains were to be returned. Sparta had resoundingly failed to destroy the Athenian empire, and in this sense Athens, whatever its financial and human losses, had won the war.

The Peace of Nicias was seen by Thucydides as an uneasy intermission between two phases of a single war. Corinth and Boeotia rejected the peace from the outset, and an energetic young Athenian politician, Alcibiades, tried to return to what may have been Themistocles' policy of stirring up trouble for Sparta inside the Peloponnese. Alcibiades' plans, like those of Themistocles, centred on Argos, once again a factor in Greek international politics after 421 and ambitious to revive mythical Dorian glories. An alliance of Athens, Argos, Elis, and Mantinea fought Sparta in 418 in the territory of Mantinea. Sparta, resolute in war as it was irresolute in politics, scored a crushing victory over its enemies. The shame of the Sphacteria surrender was wiped out in one day, and the Greek world was reminded of Spartan hoplite supremacy. If Athens, whose finances were now strong again, wanted outlets for its aggression, it would have to find them elsewhere than in the Peloponnese. It sought it first in Anatolia, second on Melos, and third in Sicily.

At some point after 425, when there was a routine renewal of the Peace of Callias, Athens began an entanglement in Anatolia with the Persian satrap Pissuthnes and subsequently with his natural son Amorges; it sent mercenary help to Pissuthnes and perhaps Amorges. If this involvement began while the Archidamian War was still in progress, it was inexplicable provocation to Persia except on the assumption that Athens was too short of cash to pay these troops itself (a 1,000-talent reserve had been set aside at the beginning of the war, but there was resistance to touching this). If the entanglement began in the period of the Peace of Nicias, it was still dangerous adventurism because nobody could say how long the peace with Sparta would last.

Thucydides says nothing about this Persian entanglement in its right place, despite its long-term importance: it was, after all, Persian intervention on the Spartan side that ultimately settled the outcome of the whole war. By contrast, he says a great deal about Athens' expedition in 416 against ostensibly unoffending Melos. Although militarily trivial, the subjugation and harsh treatment of Melos certainly had moral implications, which Thucydides explores in the famous "Melian Dialogue." It shows that the Athenians, who had made one attempt on Melos in 427 under Nicias, still wanted to round off their Aegean empire irrespective of the Dorian "ancestry" of Melos. Thucydides' debate is framed in absolute terms, as if there were no question of provocation by Melos and the only issue were whether the weaker should submit to the stronger, as Melos in the end had to do. Yet there are points to be noted. First, Melos may have contributed to the Spartan war fund as early as 426. Second, Athens had assessed Melos at the high sum of 15 talents in the context of the (admittedly optimistic) general increase of 425; there was a fugitive sense in which Melos, which did not pay this exorbitant sum, could be seen as a recalcitrant subject. This, however, is not a line pursued by Thucydides' Athenians in the "Melian Dialogue". Third, some Athenian subject allies joined in coercing Melos in 416, evidence that Ionians and Aeolians could be mobilized against Dorians and perhaps even that they positively approved of all the implications of a notably ruthless action.

Melos and the "Melian Dialogue"

In 415 Athens turned to the third and most aggressive operation of the period, the great expedition against Sicily of 415–413, better known as the Sicilian disaster. The initial commanders were Alcibiades, Nicias, and Lamachus, but the expedition was weakened by the recall of Alcibiades to stand trial for impiety (he escaped and went to Sparta, which sent help to Syracuse at his suggestion). Originally conceived in perfectly acceptable terms (a force of 60 ships to help Ionians and non-Greeks against the rising power of Syracuse), the expedition as ultimately sent was too ambitious; it consisted of a huge fleet of 140 ships—100 of them Athenian—reinforced by an additional 60. Thucydides speaks impressively but unspecifically about the cost of the expedition (he does report at one point that the Syracusans had spent 2,000 talents); from an Athenian inscription one can see that in a single transaction 3,000 talents was set aside for Sicily. A major problem was cavalry: Athens sent 250 cavalrymen without horses, but mounts were secured locally in Sicily, bringing the total to 650. (Athens also sent 30 mounted archers.) This total was not bad for a state that had never been a strong cavalry power, but it was scarcely more than half of the 1,200 that Syracuse was able to field. Even Athens' early successes in the field, and there were some, were neutralized by this disparity: pursuit of the enemy by victorious Athenian infantry became a dangerous matter because of harassment by Syracusan cavalry. When the Spartan Gylippus arrived to help the Syracusans and Athens failed to wall in Syracuse, the Syracusan cavalry made the Athenian position intolerable: those who went out from their camps foraging for food often did not come back. Nicias himself was ill but was kept in post by the Athenians, a great mistake not compensated for by the arrival of first the more energetic Demosthenes and then Eurymedon. (Lamachus had been killed in action.) The final catastrophic sea battle in the Great Harbour of Syracuse was fought in cramped circumstances that did not allow the Athenian fleet enough freedom of maneuver. The expeditionary force was virtually annihilated, including its main commanders.

The blow to Athens' morale and prestige was perhaps greater than the strictly military reverse, for, with an astonishing capacity for replacement, Athens managed, after a crash building program, to achieve rough naval parity with the Peloponnesians. This was the more remarkable in view of difficulties at home. Already before the end came in Sicily, Sparta had reopened the Peloponnesian War. On the advice of Alcibiades, the Spartans had fortified Decelea (413) and, as a result, were able to occupy Attica. Athens, embarrassed economically for this and other reasons, decided to impose a 5 percent tax on shipping instead of the tribute (but the tribute was restored in 410).

Denied the use of Attica, Athens drew more heavily on Euboea for food, and this is relevant to Euboea's revolt in 411. By then, however, there had also been revolts in the eastern Aegean and in Anatolia (413–412). As regards Anatolia, another factor is relevant: the king of Persia, angered by the Amorges affair, had decided to back Sparta. Representatives of his satraps Tissaphernes and Pharnabazus, as well as ambassadors from Chios and Erythrae, invited the Spartans to carry the war across to the eastern Aegean. This Sparta did, and in some long, drawn-out diplomacy it agreed to abandon all claim to Anatolia as part of a deal for money and a fleet; the money given was hardly lavish, and the fleet did not materialize at all (perhaps, as Thucydides hints, because Tissaphernes wanted to wear down both sides, but perhaps because it was needed

for use against Egypt. There is papyrus evidence for a revolt from Persian authority at this time, 411). For Sparta's part, it is possible that its abandonment of Anatolia was not quite final; a treaty of 408 may have stipulated autonomy for the Ionian Greeks. Despite the reservations on both sides, the possibility of a joint victory of Sparta and Persia over Athens had at least been conjured briefly into existence. For the moment, however, the war went on.

The oligarchic revolution

Athens' military resilience after its defeats in Sicily was remarkable, but the political credibility of the radical democracy had been battered: the rich had lost money, the *thêtes* had lost men, all classes had lost their illusions. This was a situation ready to be exploited by intellectual activists, who disliked the democracy anyway. Thucydides gives a brilliant picture of the oligarchic revolution of 411 (the "Regime of the Four Hundred" oligarchs), but he can perhaps be criticized for not bringing out the importance of this intellectual factor, stressing instead the general atmosphere of suspicion and terror. A complete analysis of the revolution ought, however, to allow for the influence, on oligarchic leaders like Antiphon and the less extreme Theramenes, and no doubt on others, of the subversive teaching of the sophists (rhetorically adept "experts" who professed to impart their knowledge of such politically useful skills as rhetoric, usually in exchange for money). Theramenes is said to have been a pupil of the sophist Prodicus of Ceos. Thucydides mentions sophists only once, and then not in the context of 411 at all. The first impetus to the revolution was given by Alcibiades, who certainly was a product of the sophistic age. His motives, however, were selfish and short-term (he was aiming to achieve his own recall from exile), and he abandoned the oligarchs when he failed to get what he wanted. Nor had Peisander and Phrynichos, two other leading oligarchs, always been hostile to democracy. It is certain, however, that there were some who held, as a matter of sincere theoretical conviction, that there were merits in a "hoplite franchise"—that is, an undemocratic constitution in which the *thêtes* would be barred from attending the Assembly or serving as jurors). Such a view, insofar as it was elitist, would naturally be attractive to the cavalry class, and it is an appealing suggestion that the original coup d'état was staged at the deme site of Colonus precisely because of its associations with the cult of Poseidon Hippios, "Horsey" Poseidon. But distinctions between extreme and moderate factions among the oligarchy must be made: Theramenes and Cleitophon were among the moderates who sought to justify the new arrangements by reference to Solon and Cleisthenes, who were wrongly represented, at this time, as having excluded the *thêtes* from the Assembly. (Perhaps they used the slogan "ancestral constitution," but a contemporary sophist, Thrasymachus, implies that it was on everybody's lips.) However erroneous such an appeal to Solon was with regard to the facts—it is a good example of "invented tradition"—it is undoubtedly true that members of this group behaved more moderately than some of the other oligarchs (Theramenes helped to overthrow the Four Hundred).

The Law Against Unconstitutional Proposals, a democratic safeguard, was abolished, as was pay for most kinds of political office, and the old Council of Five Hundred was to be replaced by an elected Council of Four Hundred. These changes and plans did not go unopposed. Despite its losses in Sicily, there was still a fleet, at Samos, which was not at all pleased with what was happening. And the hoplites themselves, whatever theoreticians may have wished for on their behalf, were as enthusiastic for democracy as the *thêtes*. The fleet sent a message to demand that the democracy be restored, and the extreme oligarchs were overthrown in favour of a more moderate oligarchy, the regime of The Five Thousand. This regime probably denied to the *thêtes* the right of voting in the Assembly and lawcourts, though this is controversial. In any case, it lasted a mere 10 months.

Full democracy was restored in 410, and a commission was set up to codify the law: it was evidently felt that constitutional history had been abused in 411 and that the abuse had been made possible through ignorance. Codification was to prevent a recurrence; it was expected to take four months but was still incomplete after six years. A fresh start was to be made in 403.

In 410 Athens had recovered sufficiently to win a battle against the Peloponnesian fleet at Cyzicus (this was a factor in the downfall of The Five Thousand), and the Spartans may have asked for peace; the offer, however, is not mentioned by Xenophon, who now replaces Thucydides as the main source. This was a remarkable reversal of the position in 413 when a Spartan victory must have seemed in sight. Athens, however, refused to come to terms.

The return of Alcibiades

Athenian success continued with further victories in the Hellespontine region, and Alcibiades, who had played a role in these victories, was able to return from exile in 407. He magnificently led the religious procession from Athens to Eleusis, thus atoning for, or giving the lie to, his alleged impiety in 415 when he was held to have joined in profaning the Sacred Mysteries. His subordinate Antiochus, however, lost the Battle of Notium in 406, which effectively ended Alcibiades' career. Athens managed yet another victory at Arginusae in 406. But the Athenian commanders, who failed to rescue survivors, were executed in an illegal mass trial. This was folly, and so was Athens' refusal of yet another Spartan peace offer after the battle. In this, as after the Battle of Cyzicus, it followed the advice of the demagogue Cleophon.

That a combination of Persia and Sparta could win the war easily can never have been in much doubt, even after the particular failures of trust and understanding in 411. The extra factor needed to bring it about was a combination of personalities. This happened quite suddenly after 408, with the emergence to prominence of a new Spartan, Lysander, and a new and extremely young Persian, the king's son Cyrus, sent to fight on Sparta's side. The two men got on instantly (it surely helped their relationship that Persia had made concessions, if make them it did, about the autonomy of the Greek cities). The result was the victory at the Battle of Notium and then, after the Athenian refusal of the peace offer after Arginusae, a final crushing defeat of Athens at Aegospotami (405). The Athenians were starved into surrender by Lysander (404). The Long Walls were demolished, the fleet was reduced to a token 12 ships, and the empire ceased to exist. Athens was to be governed by a Spartan-imposed oligarchy, the Thirty Tyrants.

Greek civilization in the 5th century. The effect of the Persian Wars on literature and art was obvious and immediate; the wars prompted such poetry as the *Persians* of Aeschylus and the dithyramb of Pindar praising the Athenians for laying the shining foundations of liberty and such art as the Athenian dedications at Delphi or the paintings in the Painted Colonnade at Athens itself. Less direct was the effect of the Persian Wars on philosophy. It has already been noted that famous centres of philosophy, such as Elea and Abdera, owed their existence to the Persian takeover of Ionia in 546. The thinkers for which those places were famous, Parmenides of Elea and Democritus from Abdera, were, however, products of the 5th century, and the title of "school" has been claimed both for the atomists of Abdera and for the Eleatics, who argued for the unreality of all change. A number of Ionian thinkers arrived at Athens after Xerxes' invasion, perhaps because 5th-century Ionia experienced relative material poverty and was thus no longer an agreeable place or perhaps because they had escaped from the Persian army, into which they had been conscripted. This has been suggested for Anaxagoras of Clazomenae, who impressed Socrates by identifying mind as the governing power of the universe. Another 5th-century Ionian who found his way to Athens was Hippodamus of Miletus, an eccentric political theorist, who made his own clothes and was famous for a theory of town planning. However, the laying out of cities on "orthogonal," or rectilinear, principles cannot quite be his invention (though he gave his name to such "Hippodamian" plans): such layouts are already found in Italy in the Archaic period at places like Metapontum. Hippodamus, nevertheless, may have had a hand in the orderly rebuilding of the port of Piraeus after the Persian Wars and even in the new colony at Thurii in 443. (A tradition associating him with the planning of the new

Hippodamus of Miletus

city of Rhodes, almost at the end of the century, surely stretches his life span beyond belief.)

The more theoretical side of Hippodamus' political thought did not have much detectable effect on the world around him (he thought that communities should be divided into farmers, artisans, and warriors) except perhaps for his suggestion that a city of 10,000 souls, a *myriandros polis,* was the ideal size. This is the number of colonists allegedly sent out to Heraclea in Trachis by the Spartans; and the concept of the *myriandros polis* was to be very influential in the 4th century and Hellenistic period.

It has been plausibly claimed that there is a general link between the rise of a political system, namely democracy, and the self-critical speculative thinking that characterizes the Greeks in and to some extent before the 5th century. Democracy, it is held, was causally responsible for the growth of philosophy and science, in the sense that an atmosphere of rational political debate conduced to a more general insistence on argument and proof. To this it can be objected that there are already, in the Homeric poems, remarkable debates constructed on recognizable rhetorical principles. Great warriors needed to be persuasive speakers as well. But political accountability was a cardinal principle of the Ephialtic reforms at Athens in the late 460s, and it is certainly attractive to suppose that intellectual accountability was a parallel or consequent development.

A further difficulty in assessing the relationship between intellectual activities consists in the lopsided ways in which the relevant evidence has survived. First, little is extant from any centre other than Athens, and this inevitably means that a treatment of 5th-century culture tends to turn into a treatment of Athenian culture. One can note the problem but not solve it. Second, some literary genres have survived more intact than others. Attic tragedy and comedy survive in relative abundance (the tragedies of Aeschylus, Sophocles, and Euripides, and the comedies of Aristophanes). The study of philosophy before Plato is, by contrast, a matter of detective work conducted from fragments preserved by later writers, whose own faithfulness in quotation and transmission may be suspect because of their own prejudices. (Christian apologists have perhaps been too readily trusted in this matter by students of the "pre-Socratics," or predecessors and contemporaries of Socrates.)

One set of texts that does survive in bulk and is neither Athenian in origin nor the work of poets is the Hippocratic corpus of medical writings. Hippocrates was a 5th-century native of the Dorian island of Cos, but the writings that have survived are probably not his personal work. Many of them contain references to northern Greek places such as Thasos and Abdera, a reminder that intellectual activity went on outside Athens. The most striking feature of these writings, apart from the exactness of their descriptive passages, is their rhetorically conditioned polemical character. It was necessary for the practicing doctor not merely to offer the best prognosis and cure but to disparage his rivals and show by aggressive and competitive argumentation that his own approach was superior. In fact, it seems that on one specific major medical issue the "professional" doctors did not fare as well as an amateur commentator, the historian Thucydides, who in his description of the great plague was aware, as they were not, of the concepts of acquired immunity and contagion. In other words, he thought empirically and they did not. A basically competitive attitude as well as reliance on rhetoric are features of much early prose writing; for example, Hecataeus was criticized by Herodotus, who was in turn criticized by implication, though never named, by Thucydides.

Medical writings

Such shared features are a reminder that the 5th century, before the systematization of the 4th century associated with Aristotle or the organized Alexandrian scholarship of the 3rd, did not yet make clear distinctions between literary genres. A distinction between prose and verse is perhaps implied by Thucydides' distinction between "poets" and "logographers," or writers of *logoi* (tales, accounts). His own writings, however, like those of Herodotus, show an affinity to poetry, specifically to the epic poems of Homer. Indeed, an indebtedness to epic poetry is common to both the writings of Thucydides and to the Attic tragedy of the 5th century (it seems preferable to speak of shared influence of epic poetry on both the writers of tragedy and Thucydides rather than of direct influence of tragedy on Thucydides).

Greek tragedy was not itself intended as an immediate contribution to political debate, though in its exploration of issues, sometimes by means of rapid question-and-answer dialogue, its debt to rhetoric is obvious (this is particularly true of some plays by Euripides, such as the *Phoenician Women* or the *Suppliants,* but also of some by Sophocles, such as *Oedipus the King* and *Philoctetes*). It is true that sometimes the *chorēgoi,* or rich men appointed by one of the archons to finance a particular play, were themselves politicians and that this is reflected in the plays produced. (Themistocles was *chorēgos* for Phrynichos, one of whose plays caused a political storm, and Pericles paid for the *Persians* of Aeschylus.) One play with a clear contemporary resonance in its choice of the Areopagus as a subtheme, the *Eumenides* of Aeschylus (458), however, had for its *chorēgos* a man otherwise unknown; nor is it agreed whether Aeschylus was endorsing the recent reforms or voicing reservations about them. Equally, the *Suppliants* of Euripides contains much in apparent praise of democratic institutions, but it also includes some harsh words for the kind of politician that the democracy tended to produce. Euripides' associations with the sophists (the oligarchs Cleitophon and Theramenes are specifically linked to him) are another reason why it is difficult to treat his *Suppliants* as a straightforward endorsement of democracy.

The views, political or otherwise, of playwrights themselves cannot be straightforwardly inferred from what they put into the mouths of their characters. But it must be significant that the festival of the Dionysia, at which the plays were produced, was designed to reinforce civic values and ideology in various ways: war orphans featured prominently in a demonstration of hoplite solidarity, and there was some kind of parade exhibiting the tribute of the subject allies—all this taking place before the plays were actually performed. Not even this, however, entailed that the content of the plays was necessarily expected to reinforce those civic values. The opposite may even (it has been argued) be true of some plays; for example, both the *Ajax* and the *Philoctetes* of Sophocles question the ethic of military obedience, and his *Antigone* stresses the paramount claims of family in the sphere of burial at a time when the polis had made large inroads in this area. In general, however, it is hard to believe that Sophocles, who was a friend of Pericles and served as *stratēgos* and imperial treasurer, was a kind of subversive malcontent.

The choregic system is one aspect of a (for this period) very unusual institution by which individuals paid for state projects. The 5th-century Athenian economy, though it continued to draw on the silver of Laurium and was underpinned by the more recently acquired assets of an organized empire, nevertheless looked to individuals to finance both necessary projects like triremes and strictly unnecessary ones like tragedies. It is worth asking whether such distinction between necessary and unnecessary projects is too sharp: there was a sense in which the trireme, a noble achievement of human *technē* (art or craft), was an object of legitimate pride, which might have its aesthetic aspect. That, at least, is the implication of Thucydides' unforgettable account of the rivalry between the trierarchs en route to Sicily in 415. Thucydides describes the splendid flotilla, for which publicly and privately no expense had been grudged, racing from Athens as far as Aegina out of sheer pride, joy, and enthusiasm.

The liturgy system

The psychology of contributions of this sort, the so-called liturgy system, was complicated. On the one hand, the system differed from the kind of tyrannical or individual patronage the poetry of Pindar shows still existed in, for example, 5th-century Sicily or at Dorian Cyrene, which still had a hereditary monarchy (the Battiads) until the second half of the 5th century. Athenians themselves liked to think that the system was somehow anonymous and that glory was brought on the city. That assumption was true of athletic as well as cultural success: Thucydides makes Alcibiades claim the military command in Sicily because

his Olympic chariot victories have brought glory on the city. Consistent with this, Athenian victors in the Panhellenic games were given free meals in the Prytaneium (the town hall), alongside the descendants of the tyrannicides Harmodius and Aristogiton. The evidence for this is an inscription of the 430s. On the other hand, the liturgy system was exploited for individual gain. Thus Alcibiades' plea for political recognition was an individual and traditional one, recalling the 7th-century Olympic victor Cylon, who also sought political success by his attempted tyrannical coup. It was not altogether surprising that Alcibiades' contemporaries suspected that he too was aiming at tyranny. Alcibiaded, it may be felt, can be written off as an exception and an anachronism. Far less famous speakers, however, in tight situations in the lawcourts, made comparable reference to their individual expenditure on behalf of the state, one of them frankly admitting that his motive in spending more than was necessary was to take out a kind of insurance against forensic misfortune.

Individuals might pay for the equipping of triremes, or even (like Alcibiades) own their own trireme. They might even help finance buildings like the Stoa Poikile of Peisianax (a relative of Cimon). But a building program such as that undertaken after 449 called for the full resources of the imperial state. The architects commissioned, Callicrates, Ictinus, and Mnesicles, worked under the general supervision of the sculptor Phidias; most of these men had personal connections with Pericles himself and with aspects of Periclean policy (Callicrates, for example, was involved in the building of the Long Walls). The main works on the Acropolis were temples, but even the great ceremonial gateway of Mnesicles (the Propylaea) was a lavish and expensive effort, though a secular one. The financial history of these buildings can be reconstructed with the help of inscriptions, though firm evidence for the Parthenon is lacking. Nonetheless, an inscription shows that the chryselephantine (gold and ivory) cult statue of Athena by Phidias cost somewhere between 700 and 1,000 talents, and the Parthenon itself, which housed the statue, may have cost something in the same region.

From the accounts of the Erechtheum, the temple of Athena on the Acropolis (built 421–405), it is known that highly skilled slaves as well as metics (resident foreigners) participated in the work on the friezes and columns. The

Slaves

slaves, whose work on the building can hardly be distinguished from that of their free coworkers, received payment like the rest (but the money was presumably handed over to their owners). These slaves and those used as agricultural and domestic workers (*e.g.,* the occasional nurse-companions mentioned by 4th-century orators) can be placed at one end of a spectrum. At the other end are the mining slaves working in the thousands under dangerous and deplorable conditions. Their life expectancy was short. It has been held that only condemned criminals were used in the mines, but the evidence for such "condemnation to the mines" is Roman, not Classical Athenian. Slaves were thus necessary for the working of the economy in its mining and agriculture aspects, and they also provided skills for the architectural glorification of the Acropolis. It is disputed how much chattel slaves were needed as part of the infrastructure of Athenian life in that they provided the political classes, down to and including the *thētes,* with the leisure for politics and philosophy. The answer depends on population figures, which are far from certain; perhaps the total slave population approached six figures (the adult male population in 431 was 42,000). Probably many *thētes* did own slaves. Although slaves were used for military purposes only rarely, they might exceptionally have been enrolled in the fleet. Slaves were always considered a dangerous weapon of war, but they occasionally figure prominently in descriptions of political struggle within cities; for example, at Corcyra in 427 the slaves were promised freedom by both sides but went over to the democrats. One cannot adduce this as support for an interpretation of Greek politics in terms of class struggle because the democrats may simply have made the more handsome offers.

One Athenian group that can without absurdity be called an exploited productive class was the women. They were unusually restricted in their property rights even by comparison with the women in other Greek states (Spartan women are treated below). To some extent the peculiar Athenian disabilities were due to a desire on the part of the polis to ensure that estates did not become concentrated in few hands, thus undermining the democracy of smallholders. To this social and political end it was necessary that women should not inherit in their own right; an heiress was therefore obliged to marry her nearest male relative unless he found a dowry for her. The prevailing homosexual ethos of the gymnasia and of the symposium helped to reduce the cultural value attached to women and to the marriage bond.

Against all this, one has to place evidence showing that, whatever the rules, women did as a matter of fact make dedications and loans, at Athens as elsewhere, sometimes involving fairly large sums. And the orators appealed to the informal pressure of domestic female opinion; one 4th-century speaker in effect asked what the men would tell the women of their households if they acquitted a certain woman and declared that she was as worthy to hold a priesthood as they were. In fact, priesthoods were one area of public activity open to women at Athens; the priestess of Athena Nike was in some sense appointed by lot "from all the Athenian women," just like some post-Ephialtic magistrate. (Both the inscription appointing the priestess and the epitaph of the first incumbent are extant.) The Athenian priests and priestesses, however, did not have the political influence that their counterparts later had at Rome; only one anecdote attests a priestess as conscientious objector on a political issue (Theano, who refused to curse Alcibiades), and it is suspect. It is true that Athenian women had cults of their own, such as that of Artemis at Brauron, where young Athenian girls served the goddess in a ritual capacity as "little bears." Such activity, however, can be seen as merely a taming process, preparatory to marriage in the way that military initiation was preparatory to the male world of war and fighting.

Military technology

Military technology remained surprisingly static in the 5th century. The 7th century, by contrast, had witnessed rapid innovations, such as the introduction of the hoplite and the trireme, which still were the basic instruments of war in the 5th. The 4th century was to be another period of military change, although some of the new features were already discernible in the period of the Peloponnesian War (such as the more intelligent use of light-armed troops, as in the northwest and at Sphacteria in the 420s; the more extensive use of mercenaries; and the deepened right wing in the formation of the hoplite army used at Delium). But it was the development of artillery that opened an epoch, and this invention did not predate the 4th century. It was first heard of in the context of Sicilian warfare against Carthage in the time of Dionysius I of Syracuse.

THE 4TH CENTURY

To the King's Peace (386 BC). Dionysius I of Syracuse (*c.* 430–367) can be seen as a transitional figure between the 5th century and the 4th and indeed between Classical and Hellenistic Greece. His career began in 405, after the seven troubled years in Sicily that followed the Athenian surrender in 413. For most of this period there was war with Carthage and internal convulsions that Carthage was constantly seeking to exploit. Sicily was always prone to tyranny and political instability, partly because the island was threatened by potentially hostile neighbours ready to encroach and partly because there was a large population of non-Greek indigenous inhabitants such as the forces mobilized by Ducetius (see above). Polis life never struck deep enough roots, and populations tended to be mixed and were too often transplanted. Immediately after the defeat of Athens, a radical democracy was installed in Syracuse, at the instigation of an extremist called Diocles. The leader of the moderate democrats, Hermocrates, who happened to be absent, was exiled in 410. He tried to return but was killed in 407 in an attempt (his enemies said) to establish a tyranny. Dionysius, who had been one of Hermocrates' followers (and married his daughter) seized sole power in 406. His tyranny lasted until his death in 367; it was mostly taken up by warfare, fought with fluc-

tuating fortunes, against Carthage. Successes such as the capture of Motya in 397 were hard to consolidate, and none of several peace settlements was lasting. His significance lies elsewhere than in this inconclusive fighting. The first nontorsion artillery (*i.e.*, artillery using mechanical means to winch back, by means of a ratchet, a bow of unusual solidity but of a basically conventional conception) is attested from the Sicily of this period. Torsion artillery, which used the additional power of twisted substances like sinew or women's hair to act as strings for the projection of the missile and which did not need the bow element at all, was introduced in the middle of the 4th century. Torsion-powered stone-throwing machines could be huge and could batter down massive and sophisticated fortifications. Lack of torsion artillery prevented Agesilaus in the 390s from taking fortified cities rapidly and so making progress in his invasion of Anatolia; possession of it, by contrast, helped Alexander later in the century to overrun the same area with relative ease. The preliminary discovery of nontorsion artillery in Dionysius' Sicily, however, was already a notable refinement on traditional siege technique.

In other military respects Dionysius looked to the future; his was essentially a military monarchy based on loyal mercenary power. War, which included large-scale munitions manufacture, was essential to his economy. In addition to taking on the Carthaginians in Sicily, he fought Greeks in Italy, even destroying the city of Rhegium in 386. Dionysius wanted to unite Sicily and southern Italy under his personal rule, and one need look for no subtler motive than the prestige and booty accruing from it. The kind of military monarchy he established was a crucial precedent for later figures such as Jason of Thessalian Pherae or Philip II and Alexander III the Great of Macedon.

Dionysius is called archon (an ambiguous title that can mean ruler or magistrate) of Sicily in an Athenian inscription, but he was surely thought of as king or tyrant by his local subjects. In this use of titles he has been compared to the 4th-century "Spartocid" rulers of southern Russia, Leucon I and his son Satyrus II, who (as inscriptions show) called themselves archon when dealing with their Greek subjects but king when describing their authority over the native population.

Syracuse and Sparta

The Syracuse that produced Dionysius was a late 5th-century polis both in the literal sense and in features, such as appointment to office by lot, that it had adopted from the Athenians whose invasion had just been so vigorously resisted. Dionysius himself was helped to power by Sparta, the polis that above all others remained uncompromisingly "classical" in its repeated refusal, in later times, to come to terms with the victorious Macedonians. It is a striking fact, and a further betrayal of the liberation propaganda with which Sparta had entered the Peloponnesian War, that it ended it by installing at Syracuse a tyrant who was to last for four decades; this fact was not missed by the Athenian writer Isocrates. The particular Spartans sent to help Dionysius are figures of secondary importance, but it is reasonable to see behind them the hand of Lysander, who is attested as having visited Dionysius. (There is no overwhelming reason to doubt this.)

Spartan policy immediately after the Peloponnesian War looks imperialistic in the full sense: one hears of tribute and of "decarchies," or juntas of 10, imposed by Lysander, as, for example, on Samos. The government of the Thirty Tyrants, actually a Spartan-supported oligarchy, imposed at Athens is characteristic of this short phase. The seizure, by the Athenian democrat Thrasybulus, of the frontier stronghold of Phyle in northern Attica, however, created a focus for refugees, who flocked to join him. The democrats marched south, and the extreme oligarch Critias was killed in fighting in the Piraeus. Opinion at Sparta softened, and Lysander's tough policy was reversed at Athens and elsewhere (one of the Spartan kings, Pausanias, was instrumental in this, though he himself narrowly escaped condemnation at a trial held in Sparta). This episode perhaps deserves to rank as a rare instance in which moral scruple, or at least a qualm about what the rest of the Greek world might consider unacceptable, determined a foreign policy decision by Sparta. By the end of 403, democracy was restored at Athens.

Arguably, Athenian democracy was not merely restored but comprehensively rethought at this moment. As part of a general codification of the laws, now entering its second phase (see above), it was made harder for the Assembly to legislate; instead the passing of laws (or *nomoi*), with the important exception of those pertaining to foreign policy, was entrusted to special panels of sworn jurors. The Assembly henceforth passed only decrees. Pay for attendance in the Assembly was introduced at this time, and the hillside meeting place, the Pnyx, was physically remodeled, making it easier to control admission. The Council of Five Hundred also may have been tampered with, if it is right that "bouleutic quotas"—that is, the total of councillors supplied by demes—were now altered to take account of changes in settlement patterns brought about by the Peloponnesian War. The case for discontinuity has, however, not been proved. Other post-403 changes, some not strictly datable, may be mentioned here. The Assembly no longer heard treason trials after about 350; perhaps this was because jury trial was cheaper now that the Assembly was paid. (Juries also were paid, but Assembly attendances were larger.) For the same financial reason, and perhaps also in the mid-4th century, a limit was imposed on the hitherto unrestricted number of meetings of the Assembly per prytany, or council month lasting one-tenth of the year: the limit imposed was at first three meetings, though this was later increased to four. Generals received more specialist functions in the course of the century; and financial officials, especially those in charge of funds for disbursing state pay, acquired great elected power. All this tended toward efficiency and professionalism but away from democracy. There is no doubt that the Athens of the 4th century was less democratic than the Athens of the 5th.

The Corinthian War

The restored Athenian democracy may have been less democratic in certain respects than that of the 5th century, but it was no less suspicious of, and hostile to, Sparta. These feelings, along with the straightforward hankering at all social levels for the benefits of empire (a strong and well-attested motive that should be emphasized), were to be exploited by Thebans at Athens in 395 in their appeal to Athens to join in war against Sparta. This war, called the Corinthian War (395–386) because much of it took place on Corinthian territory, was fought against Sparta by a coalition of Athens (with help from Persia), Boeotia, Corinth, and Argos. Sparta eventually won the war, but only after the Persians had switched support from Athens to Sparta. In fact, the winning side was the old combination that had proved victorious in the Peloponnesian War.

The causes of the Corinthian War lie in the policies pursued by Sparta after its victory in 404. Persian participation on Athens' side needs a special explanation, which is to be found in two ultimately related sets of operations conducted by Sparta east of the Aegean. In 401 Lysander's old friend Cyrus, the younger brother of the new Persian king, Artaxerxes II (reigned 404–359), made an attempt on the throne with Spartan help. The expedition was a military failure; Cyrus was killed at the Battle of Cunaxa north of Babylon, and the Greek army had to be extricated and brought back to the Black Sea region. It became famous, however, because a participant, first as a soldier of fortune and after Cyrus' death as a commander of the Greek force, was Xenophon, who made these exploits the basis of his *Anabasis* or "March Up-country" of the Ten Thousand. Lysander's support of Cyrus provided grounds for a change of attitude toward Sparta on the part of the new Persian king. The battle, though a short-term failure, had long-term propaganda importance because it fixed in Greek minds the possibility of a better-organized "march up-country," a project that was to be preached by the Athenian orator Isocrates, planned by Philip of Macedon and realized by Alexander the Great.

Cyrus had been given help in the early stages of his revolt by some Greek cities of Anatolia. When the Persian Tissaphernes, the victor of Cunaxa, threatened reprisals against them, they appealed to Sparta, which sent out Thibron (400). This was the beginning of the second Spartan operation in Anatolia, related to the first because the Ten Thousand were eventually able to attach themselves to Thibron, having meanwhile been harried by Tissaphernes.

Thibron's expedition was followed by that of Dercyllidas (399–397), but the most ambitious of all was led by the new Spartan king, Agesilaus, in 396. At the least (and Xenophon, a great admirer of the Spartan king, attributes to him some very grand ideas indeed) Agesilaus seems to have wanted to establish a zone of rebel satraps in western Anatolia. It is therefore not surprising that in 397 the Persians began to build a new fleet to deal with the menace of a Spartan army in Asia. (Sparta's help may, however, have had some technical justification if, as is possible, there had been diplomacy in 408 that renegotiated a more favourable position for the Ionian cities than they had been left in at the end of 411.) It may have been a further irritant that Sparta was helping another anti-Persian rebel in Egypt; the fact that Egypt maintained its independence of Persia until the 340s was a serious economic loss to the Persian landowners who had been exploiting it at a distance.

The Greece that Agesilaus had left behind was uneasy under its new Spartan masters, despite the glory of Sparta's victory over the Athenian fleet at Aegospotami (405), duly commemorated at Delphi, and the personal prestige of Lysander, who may even have received at this time some kind of cult at Samos (though perhaps only after his death in 395). In fact, Sparta was not even secure in its local dominance in Laconia and Messenia: the old helot problem recurred in 399 with the attempted revolt of Cinadon. A little farther away, Sparta's former Peloponnesian and extra-Peloponnesian allies were unhappy with what they saw as alarming extensions of Spartan territorial interests, though in fact some of these were very traditional.

The rise of Thebes

One powerful Spartan enemy was Thebes, which had emerged much strengthened from the Peloponnesian War. After the expulsion of the Athenians in 446, Boeotia had reorganized itself federally; the detailed arrangements are preserved in a valuable papyrus account by the Oxyrhynchus Historian. After the destruction of Plataea in 427, Thebes took over Plataea's vote and some of its territory; this was one reason for Theban strength. Another lay in the depredations that the Thebans had been able to carry out in Attica as a result of the occupation of Decelea. When Agesilaus prepared to leave for Anatolia, he tried to sacrifice at Aulis "like Agamemnon" before the Trojan War; but the Boeotian federal magistrates stopped him. Although they had little to fear from a Spartan presence in Anatolia, hardly a normal object of Theban ambition, Theban alarm can be explained by developments nearer home.

In central Greece in the early 390s, Sparta reinforced its position at Heraclea in Trachis and had a garrison at Thessalian Pharsalus. Initially, Lysander seems to have been at the back of this northward encroachment (good evidence connects him with Thrace and the Chalcidice). Yet because this was always a direction in which Sparta expanded if given the chance, Sparta did not pull out of central Greece during Lysander's temporary eclipse after 403. From the point of view of Thebes and Corinth, there was a risk of encirclement by Sparta. Another factor making for specifically Corinthian resentment may have been Sparta's interference in Corinth's colony, Syracuse. Unlike Thebes, Corinth had emerged badly from the Peloponnesian War; its prosperous middle class had been eroded, and this made possible a remarkable turn of events: Corinth and democratic Argos, in a unique if short-lived political experiment, became fully merged at this time. Argos, for its part, never needed much excuse to act against Sparta.

By 395 then, all Sparta's enemies were ready and willing for war. The precipitating cause was a quarrel between Locris, abetted by Boeotia, and Phocis. When the Phocians appealed to Sparta, Lysander (now back in qualified favour at Sparta) invaded Boeotia. He was immediately killed at the battle of Haliartus, however, a grave military loss to Sparta. Agesilaus returned from Asia and fought two large-scale hoplite battles but could not prize the Athenian general Iphicrates out of Corinth, where for several years he established himself with mercenaries and light-armed troops. At sea, more progress was made against Sparta: Pharnabazus and the Athenian commander Conon won a decisive battle off Cnidus (southern Anatolia) in August 394. The war might well have ended at this point, especially since Sparta faced a renewed helot threat as a result of the occupation by Pharnabazus and Conon of the island of Cythera. It was this as much as anything that made Sparta offer peace terms in 392, which would have meant the final abandoning of its claims to Asia. Artaxerxes, however, had not yet forgiven the Spartans for supporting Cyrus, and the war continued. Nor was Athens yet in a mood for peace.

In the years immediately following 392, the Athenians made such nuisances of themselves in Anatolia under Thrasybulus, who revived a number of 5th-century Athenian imperial institutions, that Persia—which was anxious to end rebellions not just in Egypt but also in Cyprus—eventually realized where its true interest lay. Consequently, it changed its support to Sparta. The Spartans under Antalcidas now blockaded the Hellespont with help from Persia and Dionysius of Syracuse, and Athens was once again starved into surrender.

The King's Peace

The ensuing Peace of Antalcidas, or King's Peace, of 386 specified that Asia, including Cyprus and Clazomenae, were to belong to the king of Persia. (Ionian Clazomenae was included because Athens had interfered there and also because its status—whether it was an island or part of the mainland—was unclear. It was in fact a peninsular site. Cyprus was included because Athens had been helping the rebel king, Evagoras.) The other Greek cities great and small, including the other islands, were to be autonomous, but Athens was allowed to keep Lemnos, Imbros, and Scyros, three long-standing cleruchies. Modern argument centres on the question of whether there were additional clauses, not supplied by the main account (that of Xenophon). For instance, the Athenian navy was perhaps ordered to be broken up and the gates on the Piraeus removed, but these may have been consequences, not clauses, of the peace. The same is true of Sparta's position under the peace, which was certainly much strengthened. There is no agreement, however, that Sparta's enhanced position was officially recognized by some such description as "champion" of the peace. Argos' merger with Corinth was cancelled, and, more important (in view of the relative power of the states concerned), Thebes had to relinquish the control of Boeotia that it had been exercizing in an unrecognized but progressively real way since 446.

In Anatolia there was little immediate change—the Spartans had after all pulled out of Anatolia some years before, though an inscription (published in 1976) suggests that the Ionian cities may have clung to a precarious autonomy until 386. One difference after 386 lay in the status of possessions up to then held by various Greek islands on the mainland of Anatolia. These possessions had hitherto been anomalous enclaves of Greek control within basically satrapal Asia, but the King's Peace surely assigned them formally to Persia in general. Anatolia now became the political property of Persia and the satraps for the 50 years until Alexander's arrival. Occasional adventures, such as Greek flirtation with the Revolt of the Satraps in the 360s, do not seriously affect this generalization.

The activities of those 4th-century satraps (and of dynasts without the satrapal title but recognized by Persia) are of great interest, though documented more by inscriptions and archaeology than by written sources. The most energetic of them was the Hecatomnid dynasty of Caria, which took its name from Hecatomnus, the son of Hyssaldomus. Hecatomnus was appointed satrap of the new separate satrapy of Caria, perhaps in the mid-390s, as a counterpoise to Sparta. He ruled his pocket principality under light Persian authority until 377 and made dedications in Greek script at a number of local sites and sanctuaries. The major Hellenizing force, however, was his son Mausolus (Maussollos on the inscriptions), satrap from 377 to 353, who gave his name to the Mausoleum, the tomb he perhaps commissioned for himself. The Mausoleum itself, a creation of Greek artists and sculptors but with some barbarian features, has long been known from surviving sculptural fragments and from Greek and Latin literary descriptions. It was constructed at Halicarnassus, which,

after a move from inland Mylasa, became the Hecatomnid capital, with palace and harbour built on monarchical lines that surely owed some inspiration to Dionysius of Sicily. The importance of other sites associated with the Hecatomnid dynasty, above all that of Labranda in the hills not far from the family seat of Mylasa, would not have been guessed from the literary sources. Inscriptions placed in aggressive prominence on fine temples and templelike buildings at Labranda (and published in 1972) attest the wealth and the Hellenizing intentions of the rulers (the dedicants include Mausolus' brother and eventual successor Idrieus). They also illustrate the range of the family's diplomatic contacts (for instance with faraway Crete) and their relations with the local communities, both Greek and native Carian. For example, in a text from Labranda, a semi-Greek community called the Plataseis confers tax privileges and citizenship on a man from Cos; the grant is ratified by yet another Hecatomnid brother and satrap, Pixodarus. And a remarkable trilingual inscription in Lycian, Greek, and Aramaic (a Semitic script used for convenience in many parts of the Persian empire), found in 1973, proves the family's interests to have spread eastward into Lycia; the text illustrates the cultural, social, and religious heterogeneity of southwestern Anatolia in the period before Alexander's arrival. Hellenization was well under way before he came.

Hellenization in Anatolia

The same conclusion is compelled by such dynastic (rather than strictly satrapal) edifices as the Nereid monument from Lycia (early 4th century) or the caryatids (roof-carrying female sculpted statues) from Lycian Limyra, a place ruled by a Hellenizing prince significantly named Pericles.

Hellenization at the cultural level and tolerance of the social structures of small local places with no military muscle did not necessarily entail favouring the political interests of the Greek states to the west. In fact, Mausolus, despite a brief and cautious insurrectionary moment in the late 360s when he joined the great Revolt of the Satraps (a movement in which there was also tentative Athenian and Spartan participation), is found actively damaging Athenian interest in the Aegean in the 350s.

In 386, however, the political dividing line between Greek and Persian interests looked relatively clean, although it was usually with the help of Greek mercenaries that over the next decades Persia made its series of attempts on the recovery of Egypt, the immediate task in the sequel to the King's Peace. Unsuccessful there, Persia had better fortune in Cyprus. In Greece, Sparta's supremacy looked as militarily imposing as in 404, though with the abandonment of Asia its moral authority was much weakened.

From 386 BC to Philip II of Macedon. The autonomy guaranteed to the Greek cities by the King's Peace in 386 represented in principle an advance in interstate diplomacy; but then as now the word "autonomy" was elastic, and Sparta by its behaviour soon made clear its intention to interpret it in the way most favourable to itself. That is, it applied the old criterion of "what is best for Sparta." Its first move, in 385, was to break up the polis of Mantinea into its four constituent villages. This move was intended to dismantle the physical polis of Mantinea as well as its democracy; in the particular Mantinean context the return to the villages strengthened the political influence of the wealthy and oligarchic landowners, whose estates adjoined the villages. The "troublesome demagogues," as Xenophon calls them, were expelled. Sparta could perhaps have represented the original 5th- or possibly 6th-century Mantinean synoecism, whereby the villages had been joined into a polis, as a breach of local autonomy (that is, of the right of the separate villages to exist as political units), but it is doubtful that Sparta even bothered to formulate any such justification. It would have been too hollow a reply to the more obvious interpretation that it was simply exploiting its supremacy to infringe on the autonomy of the Mantinean polis.

Soon after, Sparta responded to an invitation, surely welcome in view of its previous northern and central Greek involvements, to interfere against the rising power of Olynthus in northern Greece. Grown populous and powerful since its synoecism in 432 at the instance of Perdiccas II of Macedon, the city had survived the military reorganization of Macedonia by Perdiccas' successor Archelaus (413–399). Now another Macedonian king, Amyntas III, who had succeeded to the Macedonian throne about 393 after a series of short, weak reigns, joined two Greek cities, Acanthus and Apollonia, in an appeal to Sparta against Olynthus. The Spartans sent Phoebidas north, but in a momentous development he was asked into Thebes en route by a pro-Spartan faction there. Without reference (naturally) to the authorities at home, Phoebidas installed a garrison on the Cadmea, the Theban acropolis (382). The occupation of the Cadmea was a famous instance of Spartan high-handedness; indeed, it produced such a revulsion of feeling that Sparta lost its leadership of Greece. Had Phoebidas' act been promptly disowned by Sparta, the damage could have been contained. King Agesilaus, however, approached the matter solely from the point of view of Spartan advantage; he once again posed the question of whether this action had been good or bad for Sparta, with the result that Phoebidas was punished with a fine but then reemployed elsewhere, and the garrison in Thebes was retained. Meanwhile (380), Olynthus was reduced.

The occupation of the Cadmea

Agesilaus, however, gave the wrong answer to his own question; the Cadmea episode meant that Sparta would no longer have things its way. When a group of Theban exiles liberated the Cadmea in 379, they were helped by Athens, though at first unofficially. Athens, whose foreign policy in the years 386–380 had been cautious in the extreme, evidently felt it could not risk Spartan reprisals for its help to Thebes without seeking moral and military support from other Greek states. It now made a series of alliances, with Chios, Byzantium, and Methymna on Lesbos, which prefigure the formation of the Second Athenian Confederacy, formally inaugurated in 378. The charter of the new confederacy was issued at the beginning of 377. Athens was right to suspect Spartan anger; an attempted raid on the Piraeus by the Spartan Sphodrias at this time is best seen as a response to the new mood in Athens. The raid failed in its object, whatever exactly that was. Once again Sparta did not pursue the offender.

The aims of the new confederacy are set out on an inscription of cardinal importance, the "charter" document. The enemy singled out is Sparta, while the main ally is Thebes. Hostility toward Sparta, however, though it was certainly the motive shared by Athens and Thebes, does not adequately explain the participation of islanders such as the Rhodians and Chians. In these islands the main fear must have been of encroachment by such Persian satraps as the energetic Mausolus. In this respect the new alliance recalls the early 470s, when alarm felt in eastern Aegean waters about Persia's intentions had led to the formation of the old Delian League. Yet the charter says nothing about Persia or the satraps in so many words; that would have been too provocative given Athens' naval weakness at the time. On the contrary, it is likely that a clause actually spelled out an intention to remain within the structure of the King's Peace. But this is not quite certain because the relevant lines were subsequently erased, probably in a moment of Panhellenist ardour.

Action against Persia may, then, have been once again envisaged, but in other respects the precedent of the Delian League was explicitly avoided. There was to be freedom and autonomy for all as well as an allied chamber, or *synedrion,* that could put motions directly before the Athenian Assembly. An inscription from 372 shows that this chamber had an allied president. In other words, an improvement was intended on the old synod of the Delian League, which met (presumably) only when Athens called it and had no way of influencing policy in an immediate or effective way; for instance, there is no sign of allied influence in Thucydides' detailed account of the preliminaries to the great Peloponnesian War. The *synedrion* was to decide on the membership of the confederacy, and it had some financial competence. There was to be joint judicial action: although there would not exactly be a joint court, the *synedrion* was to participate in treason trials alongside the Athenian Assembly and Council.

The *synedrion*

The restrictive policies adopted by Athens are interest-

ing as showing awareness of what had been 5th-century grievances. There was to be no tribute, no governors, no garrisons, and no cleruchies. Land outside Attica was not to be cultivated by Athenians, and "unfavourable stelae" (inscribed pillars) were to be taken down. (Perhaps this is a reference to grants of the right to own land in the empire. It does not seem, however, that much or any land had survived in Athenian hands after the end of the empire in 404, and the importance of this clause may be merely symbolic.) Except for the pledge against private cultivation of land outside Attica, every one of these pledges was to be broken sooner or later, mostly sooner. There is even a hint by an orator of a private Athenian estate on the island of Peparethus, and thus one perhaps should make no exceptions at all.

Athens now began to reorganize its public finances and to build ships. A new system of levying taxes by taxation groups, Symmories, was introduced. To make sure there were no cash-flow difficulties, rich individuals were expected to produce money for the state from their own resources and then recoup it from their taxation group. The new Athenian navy defeated Sparta in the battle of Naxos (376), a victory won under the command of the Athenian Chabrias. In western waters another great Athenian commander, Timotheus, won the battle of Alyzia. These successes produced new members for the confederacy (some states had cautiously stood aloof at first). In Boeotia, which Sparta, under King Agesilaus and initially the other king, Cleombrotus I, repeatedly invaded in the years after the liberation, there was a surprising land defeat of some Spartan contingents at the hands of the Theban "Sacred Band," a crack professional force. This, the battle of Tegyra (375), anticipated the more famous Spartan defeat at Leuctra four years later. The very existence of a Sacred Band was militarily significant, indicating that Spartan professionalism was now being copied by others who would soon overtake Sparta.

By 375 these efforts had exhausted all parties, and they were ready to make peace, or rather to accept another King's Peace. (Greeks felt uncomfortable about their involvement in this kind of Persian-inspired diplomacy, in which the various peaces were "sent down"—*i.e.*, imposed—by the Persian king; as a result, the Persian aspect to this and other initiatives tends to be minimized or ignored by some literary sources, notably the "Panhellenist" Xenophon.) This time Athens' improved position was acknowledged in a clause specifically giving it the leadership by sea.

After its expulsion from Thebes, Sparta had steadily lost ground in central Greece. The Thebans energetically centralized Boeotia under their own leadership; for instance, they gained control of Thespiae and—yet again—of the unfortunate Plataea, which must have been resettled at some point, or perhaps just gradually, after the Peloponnesian War. In addition, a new power arose in Thessaly, that of Jason of Pherae, an ally of Thebes and until his assassination in 370 a military despot on the Dionysius model. Sparta was unable to respond to local Thessalian appeals against Jason, proof that Spartan ambition in central Greece had finally come to an end.

The effect of Theban expansion

Theban expansionism was bound to drive Athens and Sparta together before long. Despite renewed fighting between Athens and Sparta in the west (374 and 373) and despite Thebes' continued, though increasingly reluctant, contributions to the Athenian navy (373), it was becoming clear that Thebes was the real threat to both Athens and Sparta. In this respect the Second Athenian Confederacy, with its political justification in terms of anti-Spartan sentiment, had already been superseded by events. There were other causes for concern within the confederacy. Tribute by another name had been levied for the western operations of 373, not altogether unreasonably: ships cost money, and Athens did not have great reserves, as it had in the 5th century. Perhaps more disquieting in its implications was the Athenian garrison on Cephallenia, attested by an inscription of 373; there may, however, have been special factors, and it is not known how long the garrison remained.

At a famous peace conference held at Sparta in 371 (which, in fact, resulted in another King's Peace), Sparta tried to prevent the Thebans from asserting and formalizing their local pretensions by signing on behalf of the whole of Boeotia. After a breach in the negotiations, signaled by a rhetorical duel between Agesilaus and the Theban Epaminondas, "a man famous for culture and philosophy," as his fellow Boeotian Plutarch described him half a millennium later, the Spartans invaded Boeotia. Twenty days after the peace conference, Sparta was defeated by Thebes on the field of Leuctra, the Theban commander Epaminondas showing more than cultural and philosophical qualities. This was a major and decisive battle in Greek history. Politically, it was to loosen Sparta's hold even on its Peloponnesian dependencies and to end its long subjection of Messenia; it introduced a decade of Theban prominence (which was, however, too inconclusive in its results to deserve its usual name of the "Theban hegemony"). Militarily, the battle was innovative in several ways, not only in the sheer professionalism of the Sacred Band. The left wing of the army was deepened to 50 men, in a further development of the Delium arrangement of 424. This provided a flexible "tail," or reserve force on the left that could be deployed as the course of the battle suggested. The decision about whether, when, and how to deploy it would be the general's, whose influence on the outcome of the battle was thus greater than had been usual hitherto. By placing the best troops on the left, the Thebans aimed to knock out the best Spartan troops, who were positioned opposite them, occupying the right wing in the traditional hoplite manner. Finally, by marching forward obliquely (rather than straightforwardly, as was customary), the Thebans increased the punch administered by this deepened left.

Perhaps the Spartan defeat needs no explanation other than Theban superiority. The Spartans lost about 1,000 men, 400 of them full Spartan citizens. It is disputed, however, whether manpower problems were the most serious factor in the defeat. Aristotle, on the one hand, explicitly made the connection between the defeat at Leuctra and shortage of men. There were not enough ways for talented or physically vigorous outsiders to acquire Spartan citizenship and too many ways by which full citizens could lose their status. Thus full citizens might be degraded in status for alleged cowardice in battle, or they might fall into debt through inability to pay their mess bills (these debts often resulted in the takeover of land by women, whose social and economic position was stronger at Sparta than elsewhere). In addition, the number of full citizens was reduced by unavoidable demographic disasters such as the earthquake of 465. On the other hand, it has been replied that non-Spartans (either degraded Spartans, the so-called "inferiors" like Cinadon, or citizens of the surrounding communities) might be and probably were brigaded alongside full Spartans, at least in the 4th century.

After Leuctra there was a second peace of 371, this time at Athens. It is disputed whether Sparta participated, but it is certain that the Thebans were again excluded. It is also certain that the peace included undertakings to accept "the decrees of the Athenian allies"—a possible reference to the Second Athenian Confederacy and in any case a further strengthening of Athens' position.

The decline of Sparta

Sparta's position, by contrast, now began visibly to crumble. In Arcadia, not merely did the Mantineans organize themselves into a polis once more, but Arcadia as a whole became a federal state on the initiative of a Mantinean called Lycomedes. (The capital was to be at Megalopolis, the "Great City," a new foundation made necessary by the ancient rivalry between Tegea and Mantinea.) Both these movements were obviously anti-Spartan, and the Arcadian or federation badly needed military support from some powerful quarter. The Arcadians found it at Thebes, after being rejected by Athens (if Athens had responded positively to this appeal, major Peloponnesian developments of the 360s might never have taken place).

Federal Arcadia was in origin a local growth, but there is no doubt that Theban support was crucial for its subsequent success. Theban promotion of federalism here and in central Greece is a notable political contribution, for which the evidence is largely inscriptional. Federations are

attested in this decade not just in Arcadia but north of the Gulf of Corinth, in Aetolia, an ally of Thebes since 370, and in western Locris. There was also an intriguing Boeotian federal organization of Aegean states in the 350s, complete with *synedrion* on the Athenian model. All these federations arguably betray the influence of the Thebans, who evidently sought to export the federal principle long familiar in Boeotia itself. On a skeptical view, however, the development was a natural one and merely approximately simultaneous with the period of maximum Theban power.

In 370–69 Epaminondas invaded the Peloponnese (the first of several such invasions) and weakened Sparta irreparably by refounding Messene as a physical and political polis; the "state-of-the-art" fortifications of 4th-century Messene, an artillery-conscious circuit, stretched for nearly four miles over Mount Ithome. They are the best preserved in mainland Greece except perhaps for Aegosthena at the east end of the Gulf of Corinth; in Anatolia only Heraclea on Latmus, in Mausolus' Caria, is comparable. The loss of Messene crippled Sparta economically; in particular, Sparta no longer had a helot population to provide the economic surplus necessary for its military life-style. The combined impact of Leuctra, Megalopolis, and Messene was, however, not immediately obvious; in the "Tearless Battle" of 368, Sparta still managed to win a victory over a force of Arcadians. But Sparta was no longer a leading power.

In the 360s the main focus of Greek history shifted from Sparta to the struggle between Athens and Thebes. Neither power was really strong enough to impose a definitive solution; nor were outside forces available to give either side a decisive margin of superiority in the way that Persia had allowed Sparta to prevail in the Peloponnesian War. The 360s were a period of satrapal revolts in the western half of Artaxerxes' empire, and the subjugation of Egypt continued to elude him. In effect, though there also was some Persian-sponsored inter-Greek diplomacy in this decade, there was even less threat of force behind it than usual (the King's Peace of the 380s and that of the 370s had not been backed up by Persian men or ships). Dionysius I had added his weight on the Spartan side in 386, and his troops were found operating against Thebes as late as the early 360s. After his death, however, Sicily was not a serious factor in mainland Greek politics. Dionysius' son Dionysius II did send help to Sparta, enabling it to recover control over some formerly subject communities in 365, but that was about the limit of his interference. Dionysius II ruled precariously in Syracuse and southern Italy; he recovered Syracuse only to be finally driven to exile in Corinth by the Corinthian Timoleon in the 340s. This mid-century period of Syracusan history is of interest because of Plato's involvement in the politics of the tyranny. Dion, a relative of the older Dionysius by marriage, brought Plato to Syracuse in 367 to tutor Dionysius II in science and philosophy and generally to educate him to become a constitutional king; the visit, however, was not a success.

Dionysius II

In central and northern Greece, the energetic rule of Jason (which might have given a push to the plans of his Theban allies) had ended abruptly in 370, and his eventual aims remained and remain an enigma. Macedon was the power of the future, but that was far from obvious in the 360s. After the death of Amyntas in 370, Macedon relapsed into a period of short unstable reigns, as in the 390s. Thus neither Thessaly nor Macedon was in a position to tilt the balance of power.

Thessaly and Macedon, however, were valuable prizes. Thessaly was not only enormously fertile but also had good harbours and religious influence in the Delphic amphictyony. Macedon had ship-building timber and great natural resources (though few outlets to the sea because Greek colonial poleis stood in the way). Sparta could no longer compete for these assets, but Athens and Thebes could. Not long after the peace of 371, Athens restated an old claim to Amphipolis and added a claim to the Chersonese; in 368 it sent its general Iphicrates to Amphipolis. Thebes reacted to the Athenian claims by sending its other great man of the 4th century, Pelopidas, to Thessaly and Macedon. Theban activity in these areas did not add up to much in the end (one incidental result was that the young Philip, son of Amyntas, spent a period in Thebes as a hostage. The relevance, for Philip's subsequent army reforms, of his exposure to the methods of the first military state in Greece has often been noted). It did, however, show the Greek world the scale of Theban ambitions.

By 367, affairs in Thessaly and Arcadia were temporarily stalemated, and a peace conference was held at Susa, inside the Persian empire. Pelopidas asked that Sparta be made to give up Messenia formally and (more importantly, in view of Sparta's relative impotence at this time) that Athens be requested to give up its fleet. When these proposals inevitably failed, Thebes seized the valuable border territory of Oropus, and Athens was after all obliged to accept what was probably a King's Peace (366). There was, however, no question of Athens dismantling its navy; on the contrary, its claims to the Chersonese (reachable only by sea) were recognized in exchange, it seems, for acceptance of Theban leadership of Boeotia, including Oropus.

Athens' pursuit of essentially private Athenian aims, such as control of Amphipolis and the Chersonese, cannot have pleased its allies in the confederacy. It was costly, and it was unsuccessful. (Securing the recognition of Athenian claims in theory was not the same thing as making good those claims in practice.) On the other hand, Athens, shortly after the peace of 366, did send help—a force under Timotheus—to a rebel satrap, Ariobarzanes, in the eastern Aegean; this showed a perhaps encouraging willingness to defend Greek interests against Persia, especially since Timotheus ejected a Persian garrison he found installed on Samos. This Persian garrison was a violation of Persia's side of the original King's Peace. It may seem surprising that Athens should act against Persia so soon if the peace of 366 was really a King's Peace, but the risk of reprisals just then was slight. In any case, Timotheus' somewhat contradictory instructions were to keep to the King's Peace while also helping Ariobarzanes. Timotheus' next move, however, the installation of an Athenian cleruchy on Samos, was a capital error. Timotheus' action could be technically justified: Samos was not a member of the Athenian Confederacy, and Persia had violated the King's Peace by installing its garrison; thus the cleruchy could be seen as a military response to Persian provocation in an area not covered by the rules of the charter of 377. Nonetheless, its effect on Greek opinion was damaging, and the Thebans quickly tried to exploit it.

Some naval interest on the part of Thebes can perhaps already be inferred from its designs on Thessaly, with its good harbours. After 365, however, Theban rivalry with Athens became explicit; Thebes planned a fleet of 100 triremes, lured away Athenian allies such as Rhodes and Byzantium, and induced a revolt on Ceos. This scheme was no more successful in the long run than the Thessalian entanglement, except that the Athenian loss of Byzantium seems to have been permanent; this was a serious setback for the Athenian corn supply, given Byzantium's geographically controlling position. Thebes' Aegean *synedrion* may have been founded at this time; Byzantium was certainly a member of it in the 350s.

Theban schemes against Athens

In Thessaly, Pelopidas was killed in 364 at Cynoscephalae. Although the immediate outcome of the battle was favourable for Thebes and although Thessaly was reorganized in a way that gave Thebes for the first time an absolute majority of votes on the Delphic Amphictyony, active Theban interference in Thessaly was over.

In the meantime the Arcadian federation in the Peloponnese had split in two; the Tegean party appealed for help to the Thebans (who in turn had for allies the Argives and Messenians), and the Mantineans to Athens and Sparta. The great Battle of Mantinea ("Second Mantinea," to distinguish it from the events of 418) was a technical victory for Thebes in the strictly military sense, but (as Xenophon noted) it was actually indecisive: Epaminondas' death permanently crushed Theban hopes of leadership in Greece. The peace after the battle in effect recognized the independence of the Messenians, thus settling at the diplomatic level an issue that in reality had been settled for years. The death of Agesilaus in 360 marked the end

of one era and the beginning of another, the age of Philip and Alexander.

The rise of Macedon. In 359 two new strong rulers came to the throne, Artaxerxes III of Persia and Philip II of Macedon. The last decade of the long reign of Artaxerxes II had been blighted by revolts in the western half of his empire—at first sporadic, then concerted. Already in the late 370s Datames, the governor of Cappadocia, had established his independence. Then, by the middle of the decade, Ariobarzanes of Hellespontine Phrygia went into revolt, assisted by Timotheus of Athens and Agesilaus of Sparta. The last and greatest phase of the revolt was led by Orontes, described by the sources as satrap of Mysia. (Possibly an enclave in the Troy region of Anatolia, "Mysia" could, however, also be an error for "Armenia." If so, the geographic spread of the insurrectionist satraps was still greater.) The other rebelling satraps were Mausolus of Caria (briefly) and Autophradates of Lydia. Some participation by local Greek cities in Anatolia is possible, though perhaps they merely followed the lead of their satrapal overlords; Athens and Sparta seem surreptitiously to have helped. The aims of the revolt are a matter for speculation, but it looked serious for a long moment: a second and successful Cunaxa was a possibility. (One speculation sees the affair in dynastic terms: Orontes, who was well born, presented a greater danger to Artaxerxes than local men like Mausolus, whose ambitions were by definition limited. No one would follow a native Carian in an attempt on the kingship of Persia; it is significant that Mausolus returned to his allegiance so promptly.) At the date of Artaxerxes' death in 359, the revolt was over, the traitors' cause having been ruined by treachery among themselves. Despite setbacks, Artaxerxes II and the empire had weathered the Revolt of the Satraps.

Artaxerxes III

The new king Artaxerxes III promptly ordered the satraps to dismiss their mercenary armies, thus preempting future trouble of the same sort. This was an early indication of the vigour with which he intended to rule and which was to regain Egypt for him.

In Macedon, Amyntas had eventually been succeeded by Perdiccas, the second of his sons by Eurydice. This happened in 365, after a turbulent five-year interval of two brief reigns, those of Alexander II and Ptolemy, and one intervention by a pretender, Pausanias. Perdiccas himself was killed in 359 in a catastrophic battle against the Illyrians, Macedon's permanent enemies, and his younger brother Philip, the last of Amyntas' sons by Eurydice, succeeded.

The achievements of Philip's predecessors have naturally been overshadowed by his own, just as Philip's were to be eclipsed by Alexander's. To some extent the historical injustice is beyond redress, because the literary sources gave no systematic attention to Macedon until it was obvious that the activities of its kings were to be the determining factor in Greek history. That realization came later than 359, when Philip's chances must have looked little better than those of his immediate predecessors; thus there is not even proper information about Philip's early consolidation of power.

Fortunately, Thucydides was specially interested in the north, for personal reasons, and he speaks with admiration of the way Archelaus had pulled Macedon together militarily in the last years of the 5th century. Regarding the culture, there is valuable evidence from Herodotus and from excavations, particularly those conducted in the 1970s and '80s at Macedonian Verghina. The Macedonian kings of the 5th century were sufficiently Hellenized to compete in the Olympic Games (as Herodotus attests) and at the games for Argive Hera (as proved by a dedicated prize tripod found at Verghina). The poets Euripides and Agathon both moved to Macedon at the end of that century, and so evidently did first-rate Greek artists in the course of the next, judging from the paintings discovered in the Verghina tombs. In 1983 investigators discovered, again at Verghina, an inscription in extremely beautiful Greek lettering recording a dedication by Philip's mother, "Eurydice daughter of Sirras," which is further proof of the Hellenism of Macedon in this period.

Cultural Hellenization, however, was compatible with a



Ivory portrait head identified as Philip, c. 350–325 BC, from a tomb at Verghina (enlarged). In the Archaeological Museum, Thessaloniki, Greece.
Courtesy of M. Andronicos

social and military structure that was alien to Greek tradition, resembling instead the feudalism of later societies. (In some respects the contemporary society having most in common with Macedon was Achaemenid Persia.) The 4th century Macedonian kings made grants of land in exchange for military service; this system is hinted at by literary sources and illustrated by inscriptions. Given the size and fertility of the areas controlled by the Macedonian kings, there was huge potential for military achievement, provided Macedon's chronic enemies and invaders could be appeased or crushed.

Philip needed to buy time by means of the first method, appeasement, in order to build the army that would enable him to crush where appeasement failed. (Philip always preferred diplomacy to force, dissimilar in this respect to his son Alexander, whose preferences were the reverse.) Although Philip must have seemed unlucky in coming to the throne at so unpromising a moment in Macedonian history, there were in fact compensations, especially if one looks beyond such real but local enemies as the Illyrians and assumes that from the outset Philip's vision rested on the far horizon. The greatest hoplite power in Greece, namely Sparta, was preoccupied with regaining Messenia, just as Persia was preoccupied with Egypt. Thebes had lost Epaminondas and was soon to overextend itself badly in the Third Sacred War. Athens still had a naval empire of sorts, but this was already showing signs of breakup; in any case, if Philip was to be stopped, it would not be by sea. He could and arguably did time his operations so as to make it impossible for a fleet to get at him (ships could not sail north when the Etesian winds were blowing). On the positive side, the productivity of the silver and gold mines of the Pangaion region would be a huge asset to Philip, and thus it was encouraging that they were currently controlled by a dwarf among imperial powers, Thasos. Although Thasos seems to have been extending its mainland interests remarkably in the 360s, it was not Athens and could be dealt with.

The beginning of Philip's reign

First Philip needed to reorganize his army, which he accomplished by introducing more rigorous training and employing mercenaries. This enabled him to inflict defeats on the Illyrians and other northern enemies. At the same time he made a string of advantageous "marriages," some more official than others and scarcely amounting to more than politically slanted concubinage; one of these was to an Illyrian princess, Audata. In 357, however, all of these were effectively displaced by his marriage to the formidable Olympias, who on or about July 20, 356, gave birth to Alexander. In 358 Philip made a preliminary visit to the strategically and politically crucial area of Thessaly. He was now poised for a "blitzkrieg" against Amphipolis, which he besieged and captured in 357. Then he moved on to conquer Pydna and the mining city of Crenides, renamed Philippi (356). In 356 he formed an alliance with the Olynthians, who had good reason to be alarmed at Philip's dazzlingly rapid progress, which continued with the taking of Potidaea in 356 and the successful siege of Methone (355–354). An inscription shows that the Olyn-

thian alliance was recommended by the Delphic oracle, interesting evidence that the oracle was still politically active. The Olynthian alliance is a reminder that Philip was always happy to operate diplomatically if at all possible; in fact, the Athenians had been kept quiet at the time of Philip's assault on Amphipolis by promises that he would hand it over to them. He never did. The territory of Amphipolis was distributed to Macedonian feoffees.

After the conquest of Methone came some successes in Thrace, which Athens was unable to prevent despite attempts, a little halfhearted and a little late, to strengthen the independent Thracian princes through alliances with itself. Even the great Athenian orator and statesman Demosthenes (384–322) was slow to realize that Athens' interest required a united, not a divided, Thrace.

The "Social War"

Athens had difficulties of its own at this time. In 357 the "Social War," the war against its allies, broke out. Already in the 360s in the aftermath of the Samian cleruchy (see above), trouble had occurred on Ceos and elsewhere. In addition, Mausolus of Caria, once more loyal to Persia and its new king Artaxerxes III, and surely remembering Epaminondas' example, incited Rhodes, Chios, and Byzantium to revolt against Athens (though, as stated, Byzantium was probably already detached). Dislike of Athens was as much a factor in the outbreak of war as the intriguing of Mausolus, which Demosthenes (naturally) stressed in his search for an outside scapegoat. (Mausolus' help, however, is a fact and should not be doubted.) To Athens' costly obsession with Amphipolis and the Chersonese should be added its various breaches of the promises made in 377. (For instance, Athens had, despite the charter, installed garrisons and cleruchies and had even levied tribute under the euphemistic name of "contributions.") In fact, it did not even respect its most basic political guarantees: at the end of the 360s, the Athenian commander Chares actually helped an oligarchy to power on Corcyra.

The war went badly for Athens, and it was forced to accept a disadvantageous peace in 355 when the Persian king threatened to intervene on the rebel side. It is disputed how far the inefficiency of the Athenian navy was responsible for the defeat. There are plenty of complaints by contemporary orators to the effect that the trierarchic system was not working properly. Still, there was no absolute shortage of ships, and it has been pointed out that some features denounced by orators, such as the hiring out of trierarchic obligations to third parties, actually tended to promote professionalism, because such hired trierarchs built up expertise.

In 353 Philip was in undisputed control of a much-enlarged Macedon. He was brought into Greece itself as a result of the Third Sacred War of 355–346. This war originated in a more or less gratuitous Theban attack on Phocis, which in 362 had refused to send a contingent for the Mantinea campaign. The time lag is to be explained in terms of power politics: the Thebans had suffered a reverse on Euboea in 357, when Theban ascendancy was suddenly and humiliatingly replaced by Athenian, and they were looking for a victim. Phocian behaviour offered an excuse. The Thebans, who since 364 had influence over the preponderance of votes in the Delphic Amphictyony, persuaded it to condemn Phocis (autumn 357) to a huge fine for the usual technical offense, "cultivation of sacred land." The hope was that if, or rather when, Phocis was unable to pay, Thebes would be awarded the conduct of the ensuing Sacred War. It all went wrong. The Phocians seized the temple treasure in 356 and recruited a mercenary force of such size and efficiency that the Thebans could not defeat them. The Phocian leaders were Philomelus, followed by Onomarchus, Phayllus, and finally Phalaecus. The actual declaration of the Sacred War was delayed until 355, partly because it was only in that year that the relative impotence of one of Phocis' hitherto most impressive-looking allies, the Athenians, was revealed by the miserable end to the Social War in the Aegean.

After Philomelus' death, Onomarchus formed alliances with the rulers of the Thessalian city of Pherae. Thessaly as a whole had been willing enough to declare war on Phocis in keeping with an enmity of immemorial antiquity already remarked on as long-standing by Herodotus in the context of the Persian Wars. Nonetheless, Thessalian unity on the one hand and Theban ability to influence events in Thessaly on the other were both less than complete, and Onomarchus evidently succeeded in exploiting this fluid situation. Yet another city, Larissa, responded by issuing an invitation that was ultimately to be disastrous to Greek, as well as merely to Thessalian, freedom. It called in Philip.

The immediate consequence, a victory for Onomarchus' Phocians over Philip, his only defeat in the field, was totally unexpected. The Phocians seem to have had a "secret weapon," in the form of nontorsion artillery. In the following year (352) this defeat was, however, completely reversed at the Battle of the Crocus Field. Philip, who had already perhaps been officially recognized as ruler of Thessaly before the Crocus Field, now took over Thessaly in the full sense, acquiring its ports and its revenues. A further asset was the Thessalian cavalry, which was used to augment Macedon's own "companion cavalry" in the great battles of Alexander's early years in Asia.

The Battle of the Crocus Field

Southern Thessaly was the gateway to Greece proper, as Thermopylae had illustrated in 480 and the Spartans had recognized by their foundation of Heraclea in Trachis. A probe by Philip on Thermopylae itself was, however, firmly repelled by Athens. Philip could afford to wait and perhaps was obliged to do so by Thracian trouble closer to home (end of 352). When he laid siege to a place called Heraeum Teichos, Athens sent a small contingent in September 351. At some date not long before this, perhaps June 351, Demosthenes delivered his "First Philippic," a denouncement of Philip and Macedonian imperialism. He decried the Athenian moves to counter Philip as always being too little and coming too late. He also urged the creation of a task force and larger emergency force. It is not clear how influential Demosthenes' advice was—or how influential, at this stage, it deserved to be: at about the same time, and perhaps actually after the "First Philippic," Demosthenes was found advocating, in the "Speech on the Freedom of the Rhodians," a foolish diversion of resources to the southeastern Aegean against the encroachments of Mausolus' family. The situation there was, in fact, beyond repair.

In summer 349, with Etesian winds about to blow, Philip, despite the alliance of 356, attacked Olynthus, the centre of the Chalcidic Confederation. Olynthus turned to the only and obvious place for help, Athens. This was the occasion of the three "Olynthiac Orations" of Demosthenes. One of Demosthenes' pleas was to make the reserves of the so-called Festival, or *Theoric*, Fund immediately available for military purposes—in fact, to finance an Olynthian expedition. There is no agreement that his stirring patriotism was correct from the point of view of policy; perhaps the decision to build up Athens' financial resources slowly in preparation for the time when Philip had to be confronted nearer home was right. This unglamorous, though not actually dishonourable, policy is associated with the name of Eubulus, the Athenian leader of the pacifist party, whose caution helped to make possible the prosperous Athens of the time of the statesman and orator Lycurgus.

Olynthus fell in 348, despite the Athenian help that was eventually sent. Many of the inhabitants of the city were sold into slavery. Although Greek warfare always permitted this theoretically, the treatment of Olynthus was, nevertheless, shocking to Greek sentiment. In addition, there was no comfort for Athens from the events on its doorstep; Euboea, which Eubulus and his supporters agreed should always be defended, successfully revolted in 348.

At Athens, it must have seemed that there was no immediate further point in fighting, with Amphipolis and Olynthus gone: Philip, moreover, had been putting out peace feelers for some time. The Sacred War, however, brought Philip back into Greece, when desultory warfare in 347 caused the Boeotians to call him in; in alarm Phocis appealed to Athens and Sparta. The Phocian commander Phalaecus, however, unexpectedly declined to allow the Athenians and Spartans to occupy Thermopylae, and Athens was forced to make peace. This was the notorious

Peace of Philocrates

Peace of Philocrates—notorious because of the attempts by various leading Athenian orator-politicians to saddle each other with responsibility for what was in fact an inevitability.

The Phocians surrendered to Philip, who received their Amphictyonic votes. Many individual Phocian troops, branded as temple robbers, had already fled; some of them eventually joined Timoleon in Sicily. The cities of Phocis were physically destroyed and the remaining inhabitants distributed among villages. It is doubtful whether Philip ever seriously intended any other solution to the war in its Phocian dimension. Demosthenes was later to allege that Philip at one point had a different plan—namely, to crush Thebes and save the temple robbers in Phocis. This, however, would have been an implausible renunciation of a valuable weapon, the leadership of a Sacred War. Any such threats or promises can have been no more than feints.

Philip was for the moment supreme not merely in Phocis but in Greece. Athens, as its chief concession, had to abandon claims to Amphipolis formally. It also had to enter into an alliance, as well as make peace, with Philip. This raises the interesting question of whether Philip was already thinking of a grand crusade against Persia as early as 346; some of the sources make such a claim, but they may be contaminated by hindsight. He probably was considering such a move. For one thing, the idea of punishing the Persians for their sack of Athens in 480 was not prominent before 346 but was much heard of thereafter. Moreover, Philip had triumphantly ended one religious war and demonstrated his Hellenism and suitability for the leadership of the Greeks. Nothing would be more natural than that he or his propagandists should have hit on the idea of exploiting the still greater moral appeal to the Greeks of an all-out war of revenge for Persian impiety. In fact, the idea of a Macedonian spillage into Anatolia was a very old one indeed, and a natural one. About half a millennium earlier, the Phrygian kingdom of Midas, the predecessor of the Lydian dynasty of Croesus, had emerged as a result of a mass movement of peoples from Macedon. An Asiatic expedition is an idea that Philip could surely have thought of for himself: he did not need Isocrates to urge him, as he did in his pamphlet called the *Philippus* of 346, to settle the Persian empire with wandering Greeks (or resettle them: some of these wanderers must have been mercenaries rendered unemployed by Artaxerxes' demobilization edict of about 359). Information supplied by Artabazus, a satrap who had fled to the Macedonian court at some time in the late 350s, may have been helpful to Philip. Artabazus could have told Philip—and the very young Alexander perhaps—about the complex Persian system of supplies and travel vouchers for high-ranking officials, a system revealed to historians only in 1969, with the publication of the Persepolis Fortification Tablets. In addition, Philip seems to have had contacts elsewhere in western Anatolia, for instance with Hermias of Atarneus, a fascinating minor ruler at whose court Aristotle stayed. Whatever Philip's plans may have been, the Persian empire was not yet as debilitated or ripe for takeover as it was to be in the 330s; on the contrary, Persia suppressed revolts in Cyprus and Phoenicia in the mid-340s, and, the greatest success of all, in Egypt in 343.

In Athens after 346 there was a group who seemed to want war against Persia, and this entailed good relations with Philip. However, Demosthenes, who constantly worked against this policy, argued that Philip was untrustworthy; he pointed out that in the second half of the 340s Philip was a persistent peacebreaker, as, for instance, in the Peloponnese and on Euboea. In 344 Demosthenes even persuaded the Athenians to reject a proposed renegotiation of the peace terms offered by Philip in the person of an orator from Byzantium called Python.

Philip had preoccupations closer to Macedon in this period, which themselves make it unlikely that he wanted to upset the arrangements of 346—at least not yet. In 345 he had to deal with the Illyrians again, which he did at the expense of a bad leg wound. The strains of his intense military life had by now left their effects on his appearance. He must have looked older than his age, scarcely more than the mid-30s, because he had already lost an eye at Methone. (It is possible that a skull found in Macedonian Verghina bearing traces of a missile wound over the eye may in fact be the actual skull of Philip II of Macedon. This possibility encouraged the forensic reconstruction in 1983 of the entire head, by techniques used for rebuilding the features of unknown crash victims with a view to identification.) Philip's leg wound of 345 did not incapacitate him completely; in 344 he had the energy to reorganize Thessaly into its four old divisions, or "tetrachies." It helps to explain, however, why he was relatively inactive in Macedon until 342, when he made another and final move against Thrace, removing the first local recalcitrant, a ruler named Cersebleptes. From the economic as well as the political point of view, subduing the Thracian rulers was well worth the effort: the gorgeous Thracian Treasure from Rogozen in Bulgaria, discovered in 1986, consists of 165 high-grade silver and gilded vessels; one of them is inscribed "Property of Cersebleptes."

Reorganization of Thessaly

Philip attacked the Greek city of Perinthus in 340. Perinthus was helped by Byzantium and other Greek communities, including Athens, and even by the Persian satraps (which represents the first collision between the two great powers, Macedon and Persia). Despite all Philip's efforts (and artillery), Perinthus held out. In 340 an exasperated Philip declared war on Athens. He also switched his siege engines from Perinthus against Byzantium, but he made no easy headway there either. It is possible that the reason for Philip's abandonment of at least the second of these sieges was not military (siege engines were now virtually irresistible when applied to their target over time) but political. Philip's gaze was now fixed on Athens, the greater enemy and the greater prize.

The pretext for Philip's final involvement in Greece was trivial: still another (Fourth) Sacred War, declared this time against the petty city of Amphissa. Philip, its designated leader from the first, entered Greece toward the end of 339. This perilous occasion prompted Demosthenes' famous rallying call to Athens, reported by its author nearly a decade later in the speech "On the Crown." He urged sending an embassy to Thebes at this moment of danger for Greece as well as for Athens. Thebes responded magnificently, and the joint Greek army took up position at Chaeronea in Boeotia. The battle, fought in August 338, settled the political future of Greece until the second-century Roman conquest. No accurate account survives of the course of the battle, but it ended in a total victory for Philip. Tradition insists (probably rightly) on the valuable contribution of Alexander on the Macedonian left and suggests (perhaps wrongly) that Philip executed a feigned retreat. The Theban Sacred Band had simply ceased to exist. Athens was treated mildly, its prisoners being allowed to return home without ransom.

Philip's political settlement is illustrated by a speech wrongly attributed to Demosthenes and by an inscription much restored with the help of the speech. The settlement was a masterly construction, the League of Corinth (337). Philip had perhaps waited a little while for the inevitable pro-Macedonian reaction to set in inside the leading Greek cities. Only in Sparta, arrogant but powerless, was there no willingness to adjust. Philip invaded Laconia but did not interfere further than that. Thebes had to receive a garrison. Philip's overall goal was general acquiescence and cooperation in the war against Persia, which was now a certainty. In fact, he wanted an alliance, and without doubt the arrangements of 337 secured one. To this end most of the great federations of Greece were left intact; only Athens' naval confederacy was dissolved (though its cleruchy on Samos was retained) and, less certainly, the Aetolian League suppressed in a punitive measure.

The League of Corinth

Like the King's Peace and the Second Athenian Confederacy, the new league guaranteed freedom and autonomy. Unlike the Athenian organization, however, this new league put the emphasis on property rights. There were specific bans on "confiscation of property, redistribution of land, cancellation of debts, or freeing of slaves with revolutionary intent."

The real novelty of this league was the fact that it had a king at its head and garrisons at crucial places, such as

Chalcis and Corinth, to maintain the peace. The military requirements made of each state were set out in detail. Philip may have borrowed some of the features of the new arrangement, such as his politic use of titles, from precedents other than the Second Athenian Confederacy. Thus he may have absorbed a lesson about the politic use of titles in his mother's kingdom of Epirus; although it had been ruled by kings, the officials in the confederacy over which they presided were given Greek-sounding titles such as "secretary." Other examples may have been provided by Dionysius I of Syracuse and Leucon of Bosporus, who took different titles for use in different contexts (indeed, this may have suggested to Philip the expedient of avoiding royal titles when dealing with the Greeks: for them he would be "general with full powers"). In fact, the 4th century saw a thorough mixing of political categories, of which Philip's new league is a sophisticated example. A cruder example is present in a curious decree from Labranda, which begins with the words "It seemed good to Mausolus and Artemisia" (his sister and also his wife). Here, one finds combined a regular formula for a Greek city-state with a highly irregular decision-making body—namely, a Persian satrap and his incestuous wife.

Mixing of political categories, however, was unwelcome at home in Macedon. Perhaps some Macedonian soldiers, who might have preferred Athenian loot to an Athenian alliance, were puzzled about Philip's motives. Thus it may have been for the benefit of such doubters that, after planning his Asiatic war and sending an advance force under Attalus and Parmenio, Philip had himself depicted in a domestic Macedonian context (he would surely not have risked such a thing in Greece) as a "13th Olympian god." (Inscriptional evidence indicates that Philip may have received cult at Philippi, but cult for such founders was well established.) Further speculation about Philip's motive for this action, which is as remarkable in its way as anything he ever did, is unprofitable. For it was at this moment (336) that he was struck down by an assassin, whose own motives have never been ascertained.

Philip's self-deification

Alexander the Great. Unless Alexander was himself ultimately responsible for his father's assassination (an implausible view, but one already canvassed in antiquity), he cannot have foreseen the moment of his own succession to a father who, though grizzled, was in the prime of life. His reaction to the turn of events was remarkably swift and cool. Two highly placed suspects were killed immediately. Not many actual rivals had to be eliminated, however, because Alexander's succession was not in serious doubt. A son of Philip's brother Perdiccas, Amyntas, was still alive, but there was no reason for Alexander to see him as a threat; in any case, he was probably dead by 335.

Alexander and the Greeks. Alexander began his career of conquest in 335. He started with lightning campaigns against the Triballi and Illyrians, which took him across the Danube. Thebes was next: the Thebans had risen in the optimistic belief that Alexander had died in Illyria. He reached Thessaly in seven days and was in Boeotia five days later. Then followed the destruction of Thebes. The blame for this act is differently distributed in the two main literary traditions about Alexander, that of Arrian and that of the vulgate. Arrian, a Greek historian and philosopher of the 2nd century AD, relied on the works of two writers nearly contemporary with Alexander, Ptolemy (subsequently king of Egypt) and the historian Aristobulus. Arrian's tradition, which is regarded as the more "official" of the two, shifts the blame away from the Macedonians. The tradition of the vulgate, which is often fuller than that of Arrian, can be used to supplement or correct his. Although the vulgate tends toward the sensational, the greater reliability of Arrian can never be lightly assumed. For instance, on the Theban question, the vulgate more credibly puts the responsibility firmly on the Macedonians.

Soon after his accession, Alexander had been voted the leadership of the Persian expedition by the League of Corinth. He set out for Asia in the spring (334). Ancient writers sometimes speak in an implausible way of wars being planned by a father and executed by a son (such as the Macedonian king Perseus' war against Rome, allegedly planned by his father, Philip V). Alexander's invasion of Asia, however, is surely a clear case where a son does seem automatically to have taken over a great project, one which had been in the cards since the Battle of Cunaxa at the beginning of the century. Philip had created the army, the prosperity, and the human resources that enabled Alexander to embark on his Asian campaign. He left behind his general Antipater as governor of Greece, with 12,000 foot soldiers and 1,500 cavalry, while taking 40,000 foot soldiers (12,000 of them Macedonians) and more than 6,000 cavalry with him to Asia. To what extent Alexander needed to reorganize the army at the outset of the expedition is unclear. It is certain that he made changes during it; for instance, he incorporated Iranian troops to deal with special circumstances in his eastern campaigning and changed the structure of the cavalry so as to reduce the politically dangerous territorial affiliations of the individual brigades (*ilai*, squadrons of Macedonian cavalry, were replaced by hipparchies). From the first, however, he must have given thought to problems of reconnaissance and supply. Whereas Greek armies expected to live off the land to some extent, Alexander used wagons, despite a tradition that Philip had forced his soldiers to carry their own provisions and equipment. The core of the infantry was the Macedonian phalanx, armed with the long *sarissa*, or spear; the pick of the cavalry were the Companions, led by Alexander himself on the right wing. Philip's great general Parmenio commanded the Thessalian cavalry on the left. In addition, there were lighter armed troops, such as the scouts, and less-coordinated but highly effective contingents of slingers and other irregulars, usually from the parts of Greece where the concept of polis was imperfectly developed. This army was a formidable machine in the metaphorical sense. There also were literal machines—stone-throwing siege engines that could be assembled on the spot. The Thessalian siege engineers associated with Philip certainly continued into Alexander's reign and enabled him to conquer Anatolia and Phoenicia at comparatively high speed, given the fortified obstacles confronting him.

The Spartan Agesilaus may have hoped merely to construct a belt of rebel satraps, and Philip's ultimate aims are inscrutable. Alexander, however, as soon as he had crossed the Hellespont, cast his spear into Asian soil and openly declared that he laid claim to all Asia (admittedly a geographically fluid concept). At Troy he visited the tombs of the heroes Achilles and Ajax, paying them due religious honour; this was an early and emphatic statement that he saw himself and his expedition in epic, Homeric terms. The conquest of Asia (in the sense of the Persian empire) was more feasible than in 346: Artaxerxes III had died in 338–337, and the king now reigning was the much weaker Darius II (he succeeded in 336, after the brief reign of Arses, whom the trilingual inscription found at Xanthus in 1973 shows to have borne the title Artaxerxes IV).

It was in this region, at the Granicus River, that Alexander was confronted by a Persian army—not the central army of the Persian king but a very sizable force levied by the satraps from Anatolia itself. Alexander attacked in full daylight (the vulgate tradition of a "dawn attack" should probably be rejected); the Persians lined the opposite riverbank—impressively but suicidally. Alexander's victory was achieved in part by his own conspicuous example; he led the right wing with a battle cry to the god of battles. Such "heroic leadership" is, indeed, one of Alexander's main contributions to the history of generalship.

Alexander immediately appointed satraps in the parts of Anatolia thus acquired, thereby giving an early signal that he saw himself as in some sense the successor and continuator of the Achaemenid Persian kings, not merely as an outsider devoted to their overthrow. At the same time, he proclaimed democracy, restored law, and remitted tribute in the Ionian cities. This illustrates how seriously Alexander took the propaganda purpose of the war as revenge for the Persian impieties of 480: it is noticeable that the places he accorded specially favourable treatment in his passage through Anatolia often turn out to be places with a "good" record in the Ionian revolt or the Persian Wars; that is to say, they had been prominent rebels. Alexander felt no scruple about subjecting to direct satrapal rule the tracts

of territory outside the poleis. Whether the Greek cities of Anatolia joined the League of Corinth is an intractable question. Some of the islanders certainly did, as, for instance, Chios, where an inscription recording the terms of Alexander's settlement proclaims bluntly that "the constitution is to be a democracy" and refers to the "decrees of the Greeks." As for Asiatic cities like Priene, there is no certainty, but the probability is that they joined the league.

Priene was a very old city indeed, one of the Ionian "Dodecapolis," but it was physically derelict. It is possible that Alexander in some sense refounded this and other western Anatolian Greek cities, such as Heraclea south of Latmus and Smyrna. (There is, however, an almost equally strong case for associating their physical reconstruction with the Carian Hecatomnids, the family of Mausolus). If Alexander was their founder, this would be the first good evidence of the urbanizing that was a marked feature of his policy for the conquered territories to the south and east. In this respect, however, as in others, credit should be given to Philip for his example: Philippi (the renamed Crenides) was not his only city foundation.

The siege of Halicarnassus

At Halicarnassus, Alexander met his most serious resistance so far from a defended city, in mid-334; Miletus had not delayed him long (nor was it punished very severely—it had after all been the leader of the Ionian revolt). The siege of Halicarnassus was a far tougher operation. The city had good defenses, both natural and artificial, and had been chosen as the local Persian military headquarters. The fighting was severe, though in the apologetic tradition used by Arrian the severity is minimized. At one moment Alexander was forced to the extremity of having to send a herald to ask for the bodies of some Macedonians who had fallen in front of the walls. After the city was taken—the citadels held out for another year or two—Alexander reappointed the native princess Ada as satrap (his earlier satrapal appointees had been Macedonians). She was the sister of the great Mausolus, and her reinstatement prefigures Alexander's shrewd subsequent policy of allowing local men and women to remain in post (though usually, like Ada herself, under the superintendence of a Macedonian troop commander). A romantic story makes her "adopt" Alexander as her son, a gesture graciously accepted by Alexander. That gesture of conciliation toward the native population was good politics.

After the conquest of Halicarnassus, Alexander moved east, meanwhile sending to Macedon for drafts of reinforcements. The scale of these demands through the whole campaign and their effects on the domestic situation in Macedon are not easy to estimate; the record of the literary sources is too fitful and episodic. According to one view, Alexander's legacy was one of lasting damage; he had exhausted the manpower of Macedon to such a point that the Macedon of Philip V and Perseus inevitably succumbed to the Romans with their almost infinite capacity for replacement. On the other hand, one must allow in the reckoning for a good deal of voluntary emigration by Macedonians to the armies and cities of the successor kingdoms in the Hellenistic period. Thus Alexander was not the only culprit; there were more intangible demographic forces at work.

Alexander's path took him from Carian Halicarnassus to Lycia and Pamphylia. At about the Lycian-Pamphylian border a strange natural phenomenon occurred that allowed Alexander and those with him to enjoy a freak dry passage along the coastline. This was greeted by his supporters as a portent and a recognition of Alexander's divinity (the sea "doing obeisance" to the great man). It was the first believable suggestion that special religious status could be claimed for Alexander.

In early 333 Alexander moved through Pisidia, where the nearly impregnable mountain city of Termessus, a remarkably well-preserved site 21 miles northwest of the modern Antalya, managed to hold out (even Alexander's early years in Asia were not an uninterrupted success story). Morale and self-esteem had to be satisfied with the taking of Sagalassus and some minor places. Thus it was high time for a piece of propaganda and political theatre, especially since the Aegean he had left behind him was not altogether quiet. A Persian counteroffensive was achieving some notable reconquests (but eventually troop drafts were required by Darius for the campaign that finally took shape at Issus, and the Aegean war shriveled to nothing).

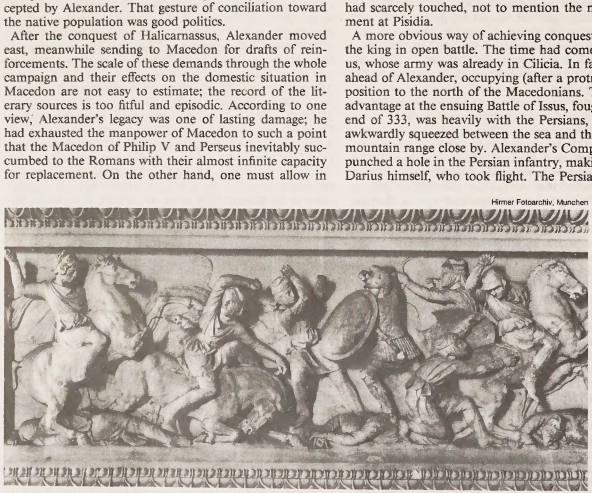
The Gordian knot

Alexander found his opportunity for propaganda some distance farther north in the Anatolian interior at Gordium, the old capital of the Phrygian kings (themselves, as stated, ultimately of Macedonian origin). There occurred the famous episode of the "cutting of the Gordian knot." The old prophecy was that whoever unloosed the knot or fastening of an ancient chariot would rule Asia. Alexander cut it instead—or perhaps pulled out the pole pin, as one tradition insisted. Either way, he solved the problem by abolishing it.

The visits to Pisidia and Phrygia had been a huge detour, evidently designed to show that Alexander had conquered Anatolia. This statement raises problems of definition; "conquest" was a relative term when there were large tracts of Anatolia, such as Cappadocia, that Alexander had scarcely touched, not to mention the mixed achievement at Pisidia.

A more obvious way of achieving conquest was to defeat the king in open battle. The time had come to face Darius, whose army was already in Cilicia. In fact, Darius got ahead of Alexander, occupying (after a protracted delay) a position to the north of the Macedonians. The numerical advantage at the ensuing Battle of Issus, fought toward the end of 333, was heavily with the Persians, but they were awkwardly squeezed between the sea and the foothills of a mountain range close by. Alexander's Companion cavalry punched a hole in the Persian infantry, making straight for Darius himself, who took flight. The Persian mercenaries

Hirmer Fotoarchiv, Munchen



Alexander in battle, detail from the so-called Alexander Sarcophagus, marble, *c.* 310 BC, from Sidon. In the Archaeological Museums of Istanbul.



Alexander's empire at its greatest extent.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte*, Georg Westermann Verlag, Braunschweig

were routed by the Macedonian phalanx. After the battle, Darius' wife and mother both fell into Alexander's hands. In an exchange of letters Alexander grandly offered that Darius could have them back—"mother, wife, children, whatever you like"—if he recognized his own claim to be lord of Asia and addressed him as such for the future. Darius, of course, refused the offer.

Alexander did not immediately follow Darius eastward; instead he continued southward in the direction of Phoenicia and eventually Egypt. The Phoenician cities of Byblos and Sidon submitted willingly, but Tyre was a major obstacle. Its walls were not finally breached until summer 332, after various contrivances had been tried, including a huge and elaborate siege mole. The siege of Gaza occupied much of the autumn; when the city at last surrendered, Alexander dishonoured the corpse of Batis, its commander, in the way that Achilles in the *Iliad* had treated the corpse of Hector. Alexander's imitation of Homeric heroes had its less attractive side.

This was the part of the world in which the Jews might have encountered Alexander. No doubt there was some contact, but virtually all the available evidence is unreliable and romantic or even fabricated to give substance to later Jewish claims to political privileges. Alexander's effect on the Jews was indirect, but no less important for that: he surrounded them with a Greek-speaking world.

Egypt was taken without a struggle, an indication of the dislike the subject population felt toward Persia. (Even though Egypt had been reconquered by Persia hardly more than a decade before, it is possible that there had been yet another revolt since 343.) Alexander's period in Egypt was marked by two major events, the founding of Alexandria and the visit to the oracle of the god Ammon at Siwah in the Western Desert. Although the sources disagree about which event came first, the foundation probably preceded the visit to the oracle.

Alexander in Egypt

The new city of Alexandria, the first as well as the most famous and successful of many new Alexandrias, was formed by joining a number of Egyptian villages (April 331). Alexander supervised the religious ceremonies of foundation, including Greek-style athletic and musical games (an indication of his intentions to Hellenize these foundations, at least as far as their cultural life was concerned); he thought that the site was an excellent one and hoped for its commercial prosperity. It is quite certain, from an inscription, that early Hellenistic Alexandria possessed a civic council; this and other self-governing institutions such as an assembly probably go back to Alexander's time. Not all Alexander's foundations were run on this liberal model, though some were inaugurated with similar symbolic gestures in the direction of Hellenism. One hears of "satraps and generals of the newly founded cities," a phrase that does not imply much self-government. No doubt some of Alexander's "new foundations" were little more than military camps; and one should assume that in all the far eastern Alexandrias the native population was forced to perform menial or agricultural tasks.

The oracle of Ammon at Siwah, to which Alexander now made a pilgrimage, was already well known in the Greek world. Pindar had equated Ammon with Zeus, the oracle had been consulted by Croesus in the 6th century and Lysander in the 5th, and there was a sanctuary to Ammon at Athens in the first half of the 4th. Alexander had a *pothos,* or yearning, to visit Ammon (the word *pothos* is often used by the sources to describe his motives and is appropriately suggestive of far horizons, even if it does not reflect a usage of Alexander's own). He wanted to find out more about his own divinity, the implication being that he already had an inkling of it. He was told what he wanted to hear; more than that (some sources offer a great deal more) was probably speculation to fill a gap.

Alexander then crossed Phoenicia again to meet Darius for the second and last time in the open field at Gaugamela (between Nineveh and Arbela) at the beginning of October 331. The tactic was to be the usual one—a leftward charge by Alexander from the right wing toward the centre, while Parmenio held the left wing firm. Parmenio seems, however, to have encountered unusual difficulties and had to summon help from Alexander, who was already in victorious pursuit of Darius. The mechanics of this "summons" are not clear, and the story may be a fabrication intended to discredit Parmenio. Alexander and his troops won the battle, sealing the fate of the Persian empire, but Darius managed to escape. Alexander then moved to Babylon, where in another gesture of conciliation toward the Iranian ruling class he reappointed Mazaeus as satrap, with Macedonians to supervise the garrison and the finances.

This kind of gesture has been much discussed; it can

be both overinterpreted and unduly minimized. Ideas that Alexander, then or ever, planned to forge a harmony between nations at a mystical level have no solid basis in the evidence. There is nothing odd, however, in supposing that his intentions toward Persians like Mazaeus (or Abulites, confirmed in the Susa satrapy about the same time) were positive. Also, the idea of such "fusion" was not entirely new. Greeks such as Xenophon earlier in the century had by their writings and actions anticipated Alexander's policy of fusion, and the cooperation of Cyrus and Lysander was just the most famous example of mutual understanding between Persian and Greek. Nor is it convincing to interpret Alexander's policy of integrating army and satrapy as a repressive device. Macedonians like Peucestas (appointed satrap of Persia) learned Persian and were rewarded for it; and Hephaestion's position of favour with Alexander is largely to be explained by his support of Alexander's Orientalizing policies. Admittedly, after leaving Iranian territories, Alexander returned to employing Macedonians as, for instance, in the Indus lands; but even there one finds native appointees like the Indian king Porus. Military integration—the use of Iranian horse-javelin men—is first firmly attested soon after the battle of Gaugamela. This is to be explained in purely military terms: the Companion cavalry on their own were not entirely suited to the more disorganized warfare lying ahead against the fierce opponents waiting to the east and north of Iran proper.

Alexander's policy of fusion

After some spectacular campaigning in Persis proper, Alexander occupied the palace of Persepolis, where the strong defensive position known as the "Persian Gates" was taken only after an unsuccessful and costly initial assault. The palace of Persepolis was looted and burned (spring 330). The less creditable tradition of the vulgate maintains that the fire started when a drunken Athenian courtesan called Thais led a revel that got out of hand, and this may well be right. The event, however, could be exploited afterward as a signal to dissident Greeks at home that the "war of revenge" was complete.

To securely establish the propaganda value of the burning of Persepolis would require a more precise chronology for the phases of that Greek dissidence than is ever likely to be achieved. The last fling of 4th-century Sparta was a revolt led by its king Agis III. It was probably still going on in 330 when it culminated in a narrow victory by the Macedonian general Antipater over the Spartans at Megalopolis. If this is right, the burning of Persepolis at about this time makes good propaganda sense. Athens had not participated in the revolt. The quiescence of Athens in the early years of Alexander's campaigning is to be explained partly by the policy of civic retrenchment associated with the name of Lycurgus (a phase of Athenian history that included a remarkable building program, the first since the 5th century) and partly by a well-attested grain shortage in Greece, which may have sapped the will to fight.

The conquest of Bactria and the Indus valley. By the middle of 330 Darius had been killed—not by Alexander but by his own entourage. Alexander now adopted symbolic features of Persian royal dress, but one of Darius' noble followers (and murderers), Bessus, the satrap of Bactria, also proclaimed himself king. The reckoning with Bessus, however, had to be postponed until the middle of 329. Alexander, who had initially followed Darius north, now moved steadily east, through Hyrcania and Areia, where Satibarzanes was confirmed as satrap; Alexander planned an invasion of Bactria and the elimination of Bessus. Satibarzanes, however, revolted almost instantly, and Alexander turned south again to deal with this rebellion. Having done so (though without taking the satrap himself), he maintained direction southward, toward Arachosia and Drangiana, home satrapies of Barsaentes, another of Darius' murderers. Barsaentes, however, fled to India.

At Phrada, capital of Drangiana, occurred the most famous conspiracy of Alexander's expedition, that of Philotas, the son of Parmenio and a commander of the Companion cavalry. There was little solid evidence for the prosecution to go on, but it is clear that Alexander's Orientalizing tendencies and the ever more personal style of Alexander's kingship had begun to irk his Macedonian nobility, accustomed as they were to express themselves freely, as in the outspoken court of Philip's day. Philotas had no doubt spoken very incautiously on some sensitive subjects, such as Alexander's visit to Ammon. The execution of Philotas entailed the execution of his father Parmenio as well, not because there was any serious suggestion that he too had been plotting but as a matter of practical politics; the family group of Parmenio, which can be elucidated by means of prosopography (the investigation of family ties with the help of proper names), had considerable power.

Execution of Philotas and Parmenio

The year 329 saw the final elimination of Satibarzanes and the capture of Bessus in Sogdiana, north of the Oxus River from Bactria. In Sogdiana, Alexander founded the city of Alexandreschate, "Alexandria the Farthest," not far from the site of Cyropolis, a city of Cyrus II the Great, whom Alexander highly admired. This is a reminder that Persian urbanization in Central Asia had not been negligible. (At the interesting Bactrian site of Ai Khanum, which cannot definitely be identified as an Alexandria, there is evidence of Achaemenid irrigation.) Alexandreschate was a prestige foundation, designed, as explicitly stated by Arrian, for both military and commercial success. Alexander had already planted a number of new Alexandrias in central Iran, including Alexandria in Areia (Herat), Alexandria in Arachosia, and almost certainly Qandahār, on the exciting evidence of a metrical inscription found there by a British excavation team in 1978. There was another major foundation called Alexandria in the Caucasus at an important junction of communications in the Hindu Kush. How far Alexander intended these places to be permanent pockets of Hellenism is not clear. That Hellenism could survive in these regions is shown by the case of Ai Khanum, which had many of the features of a Greek polis, including gymnasia and an agora with an *oikist* (city-founder) cult; there are even inscribed Delphic religious precepts. Nonetheless, many of Alexander's Greek colonists in Bactria tried to return home to the mainland immediately after his death out of *pothos,* or yearning for their Greek way of life.

Bessus and Satibarzanes were not the last satraps of eastern Iran to offer resistance. It took fully two years (until spring 327) to suppress Spitamenes of Sogdia and other tribal leaders. The period was full of strain, culminating in the disastrous quarrel between Alexander and Cleitus, one of his senior commanders and the newly appointed satrap of Bactria at the end of 328. The quarrel ended in Alexander actually killing Cleitus with his own hands in drunken fury. The issue was a personal one, which, however, merged with a matter of principle: Cleitus had criticized Alexander's leadership (there had admittedly been at least one military reverse due perhaps to inadequate planning), comparing him unfavourably with his father Philip. Before the army moved in the direction of India, there were two more incidents that widened the gap between Alexander's conduct and traditional Macedonian attitudes. First, Alexander attempted to introduce the Persian court ceremonial involving *proskynesis,* or obeisance. Just what this entailed is disputed; perhaps it amounted to different things in different contexts, ranging from an exchange of kisses to total prostration before the ruler in the way a Muslim says his prayers. What is not in doubt is that for Greeks this meant adoration of a living human being, something they considered impious as well as ridiculous. It was the court historian Callisthenes who voiced the feeling of the Greeks. The *proskynesis* experiment was not repeated: Alexander did not in the end insist on it. It is difficult, however, not to connect Callisthenes' role in this affair with his downfall not long after, allegedly for encouraging the treason of a group of royal pages. This was the second of the two alarming incidents of the period.

India was the objective in 327, though Alexander did not reach the Indus valley until 326, after passing through Swāt Cas from the district of the Kābul River. In 326, at the great Battle of the Hydaspes (Jhelum), he defeated the Indian king Porus in the first major battle in which he faced a force of elephants. How much farther east Alexan-

Conquest of India

der might have gone is a question that has fascinated posterity, but the curiosity and patience of his army was exhausted. At the Hyphasis (Beās) River he was obliged to turn back.

Alexander did not, however, retrace his path but took the route southward through the Indus valley toward the Arabian Sea and the Gulf of Oman. He narrowly avoided death at the so-called "Malli town," where an arrow seems to have entered his lung. The subsequent march westward in 325 through the desert region of Gedrosia (Balochistān) was a death march; its horrors emerge vividly enough from the literary narratives, but they are certainly understated. Alexander's motive for ordering the march may have been the desire to outdo the mythical queen Semiramis and the legendary Cyrus the Great; but the scale of the catastrophe does suggest that his judgment was by now badly impaired. Meanwhile, Nearchus led the fleet from the mouth of the Indus to that of the Tigris, a voyage recorded by Arrian in his *Indica*, using the account of Nearchus himself.

The final phase. In Carmania, to the west of Gedrosia, Alexander first staged a week-long drunken revel, in which he himself posed as Dionysus, as a release of tension after the preceding nightmare journey. Then he ordered his satraps and generals to disband their mercenary armies, like Artaxerxes III in 359 and perhaps for the same reason, namely, fear of insurrection. This was a period of punitive action against disobedient or negligent satraps. One official who in this atmosphere preferred to abscond rather than brazen out the inquisition was Harpalus, the royal treasurer, who made his way eventually to Athens. The exact fate of the money he took with him was and still is a celebrated mystery. The fact that Harpalus' activities as treasurer had evidently been quite unsupervised was typical of Alexander's short and impatient way with administrative problems. (It is most unlikely that he planned an ambitious financial restructuring of the empire, giving special responsibilities to men with the right expertise. One finds men like Cleomenes in Egypt or Philoxenus in Anatolia combining territorial with financial responsibilities, but no general conclusions can be drawn.)

At Susa in 324 Alexander staged a splendid mass marriage of Persians and Macedonians. He himself had already married a Bactrian princess, Rhoxane, in 327, but he now took two more wives, a daughter of Darius III called Barsine (or Stateira) and Parysatis, the daughter of Artaxerxes III. This and other demonstrations of "Orientalizing," including the brigading of Iranian units into the army, overcame a final mutiny at Opis near Babylon. After haranguing the troops, threatening them, and finally sulking, Alexander won back their affections; following this meretricious and emotional performance, he chose to heal the rift symbolically by a more organized piece of theatre, a great banquet of reconciliation (thus demonstrating for the last time in Archaic and Classical Greek history the usefulness of the banquet, or symposium, as an instrument of social control).

Other actions or schemes in this final phase were of the same megalomaniac type: a request for his own deification, sent to the Greek cities; a demand that they take back their exiles; a monstrous funeral pyre for his dear friend Hephaestion (never completed); and a plan of circumnavigation and conquest of Arabia. So much is well documented. Lists of other spectacular last plans survive, but they are hardly needed; the achievements of the last 13 years were extravagant enough. Alexander died at Babylon, after an illness brought on by heavy drinking, in the early evening of June 10, 323.

Greek civilization in the 4th century. The 4th century is in many ways the best-documented period of Greek history. There is, admittedly, a greater number of documents from the 3rd century, when inscriptions and papyri abound (there are virtually no documentary papyri before the time of Alexander). The writings of the 3rd-century prose historians, however, are mostly lost. In the 4th century, by contrast, there is an abundance of evidence of all kinds. Inscriptions are much more common than in the 5th century and begin to appear in quantity from states other than Athens. Forensic oratory from the 5th century has scarcely survived at all, but from the 4th century there are more than 60 speeches attributed to Demosthenes alone. Most of this corpus of oratory is set in an Athenian context, but one speech of Isocrates deals with business affairs on Aegina. Although there is no 4th-century tragedy and no epinician poetry like that of Pindar, the comedies of Aristophanes from the beginning of the century and those of Menander from toward the end have survived. These are illuminating about social life, as are the prose writings of Aristotle's pupil Theophrastus, especially his *Characters*. The writings of Plato, in their anxiety to define an ideal polis invulnerable to stasis or civil strife, give evidence of the instability of the 4th-century world in which it could be said that in every city there were two cities, that of the rich and that of the poor. Aristotle's *Politics* examines the theoretical conceptions underlying Greek attitudes toward polis life. This is a precious document, although it can be criticized for insufficient awareness of the monarchical and federal developments of the age.

The state of the sources

No such criticism can be leveled at the historiography of the age. It is from Xenophon that one learns of the grand plans of Jason of Pherae, and knowledge about Dionysius I is derived, by less direct routes, from the 4th-century historians of the Greek west Ephorus, Philistus, and (toward the end of the century) Timaeus of Tauromenium. In fact, the process of explaining history in terms of personality already begins with Thucydides, who arguably came to see that a dynamic personality like Alcibiades could by sheer charisma and force of character have an impact on events irrespective of the content of his policies. It was surely this aspect of Thucydides' work that Aristotle had in mind when he defined history as "what Alcibiades did and suffered." Aristotle's nephew Callisthenes began by recording the history of the city-states in a fairly traditional way (which, however, did more justice to the Theban hegemony than had that of Xenophon), but then he joined Alexander's staff in order to write the *Deeds of Alexander*. Evidently, history was now seen as what Alexander did and suffered. Even earlier than that, however, the central role of Philip's personality had been acknowledged by Theopompus of Chios, who (like Callisthenes) moved in the direction of writing history that revolved around the person of a king; he called his history of Greece *Philippica*. "The Affairs of Philip." Meanwhile there were local historians of Attica, such as Androtion, who continued to value Athens' past and even ventured to rewrite (not merely to reinterpret) the facts about it. These men, who are known as Atthidographers, were not simply antiquarians escaping from the monarchic present. On the contrary, the greatest of them, Philochorus, was put to death in the 3rd century by a Macedonian king for his excessive partiality toward King Ptolemy II Philadelphus of Egypt. All these authors were, in different ways, coming to terms with monarchy.

In addition to works of history there are 4th-century treatises that show how Greeks experienced the new military monarchies. Xenophon's *Cyropaedia*, or "Education of Cyrus," is a novel about Cyrus the Great, but it is also a tract on kingship and generalship addressed to the class of educated Greek commanders and would-be leaders. (In comparable fashion Isocrates offered advice on kingship to the semi-Hellenized rulers of Cyprus.) The surviving treatise on siege-craft by Aeneas of Stymphalus in Arcadia (known as Aeneas Tacticus) is valuable not only for the evidence it provides about dissensions inside a polis—there is an entire section on "plots"—but also for the awareness both of the ruthless methods of men like Dionysius, who figures prominently, and of the new military technology of the age. (The treatise includes, for example, practical advice on how to defend walls against battering rams.) Aeneas Tacticus' treatise, more than any other surviving prose work of the 4th century, makes the point that this was an age of professionalism. Many technical monographs are known to have been written in this period but have not survived. For instance, Pythius, who worked on the Mausoleum, also wrote a book about another of his projects, the Temple of Athena Polias at Priene. (There were 5th-century precedents for some of this: Polyclitus of Argos had written a famous treatise on proportion in sculpture and Sophocles a monograph about the chorus.)

Technical monographs

In the sphere of architecture, the 4th century produced

no Parthenon, but it was the great age of military structures. Most of what survives of the elegant fortifications of the northwestern frontier demes of Attica stems from the 4th century; inscriptions attest refurbishing work on Phyle in particular at about the time of the Battle of Cheronea. Outside Athens there were big projects, such as the temple at Epidaurus and the Mausoleum at Halicarnassus.

Buildings such as the Mausoleum were commissioned by powerful individuals, further proof that the emergence of commanding personalities is a noticeable feature of the 4th century. In some respects it represents a return to Archaic values: a tyrant like Dionysius has much in common with Peisistratus of Athens or Polycrates of Samos, and Philip II of Macedon can be seen as comparable to Pheidon of Argos, a hereditary monarch who transformed his power base into a military autocracy. Revised attitudes toward such individuals are already detectable near the end of the 5th century. It seems that, when Athens founded Amphipolis in 437, its founder Hagnon, father of the oligarch Theramenes, was given some kind of cult in his lifetime. That is the usually neglected implication of a passage of Thucydides, which definitely records the award of cult honours at Amphipolis to the dead Spartan general Brasidas after 422. In the early 4th century another Spartan, Lysander, received cult at Samos, and later in the century Euphron, a tyrant at Sicyon, was buried in the agora "like a founder."

At Athens itself, before the request by Alexander for his own deification, there could be no question of divine cult for a living man (although it is possible that Alexander had already arranged some kind of hero cult at Athens for Hephaestion). Nonetheless, even at Athens there was a marked trend toward more assertive monuments. This is particularly evident in the commemorative choregic monuments built to celebrate victories in the great Athenian festivals. The most famous of these, the Choregic Monument of Lysicrates, which used to be called the "Lantern of Demosthenes," represents a transitional phase; its inscribed dedication falls between the anonymity (actually more pretended than real) of the corporatist benefactions of Classical Athens and the assertiveness of Hellenistic Greece with its emphasis on individual generosity. On the one hand, the inscription makes clear that what is celebrated is victory by the tribe as a whole; on the other, the great prominence of the man's name stresses individuality, as does the idiosyncratic form of the monument. Clearly, this is an emphatic statement in the first person singular. Consistent with these developments is the marked tendency toward portraiture in art. Persian satraps such as Tissaphernes issued coinage with what were obviously meant to be realistic depictions of the satrap's head. Individual rulers were represented by statues in the round, like that of "Mausolus" from the Mausoleum (which may or may not be an attempt to represent Mausolus himself but which incontrovertibly is a portrait of some powerful individual), or by figures on friezes, as those on the "Alexander Sarcophagus" in the Archaeological Museums of Istanbul. Although the workmanship is evidently Greek, the ethos is uncompromisingly royal. Alexander created a new visual image for himself: unlike the bearded Philip, Alexander is portrayed as clean-shaven, young, and idealized. Lysippus, in particular, is said to have caught Alexander's physical qualities in his royal sculpture portraits.

Portraiture in art

The Athenian empire had given employment to many artists, architects, and sculptors, both from Athens itself and from the subject states of the empire. When the empire collapsed in 404, many of these had to seek employment elsewhere. Some went to the courts of satraps like Mausolus or of military rulers like Dionysius: both of these had money to spend on art, building, and fortifications. Another wealthy court was that of Macedon. One remaining recourse in Athens, however, was funerary art; the most famous funerary stelae and sculptured monuments found at Kerameikós, the city's prestigious cemetery, date from this period, before such lavish commissions were outlawed by the Athenian ruler Demetrius of Phaleron after Alexander's death. Some of those buried were foreigners; for instance, there was a precinct for the Messenians, one for some immigrants from Heraclea on the Black Sea, and one for those from Sinope, also in the Black Sea region. (In the Archaeological Museum of Piraeus there is a monument comparable to another one of a Black Sea immigrant, a reminder of Athens' commercial connections with this crucial grain-growing area.) In the Kerameikós there is even a grave of a Persian with a larger-than-life torso of a seated man in Persian dress.

Whatever the political effects of the King's Peace of 386, it was evidently not a barrier to social and commercial exchanges. Inscriptions in the corpus of Demosthenes' speeches frequently mention trade with ports in Phoenicia and Anatolia and occasionally allude casually to piracy, a classic by-product of such trading activity. There is epigraphic evidence for piracy as well: in the 340s Athens honoured Cleomis, tyrant of Methymna on Lesbos, for ransoming a number of Athenians captured by pirates. Lesbos had always enjoyed trading links with the Black Sea region, and in the 4th century more than ever. One should imagine Athenians and metic Athenian traders (*i.e.,* foreigners resident at Athens) going in numbers via Lesbos and the Sea of Marmara to the rich granaries of southern Russia. Some no doubt settled in these regions, though the inscriptional evidence for Athenians abroad in the 4th century (as opposed to evidence for foreigners settling in Athens or Piraeus) is in need of systematic collation.

Immigration and free movement of individuals between one polis and another are typical features of the 4th century. They are best documented for Athens but hardly confined to it, given the attractiveness of the royal and satrapal courts. At Athens itself, the great magnet for immigrants was naturally Piraeus, the city's densely populated, multilingual, multiracial port. Bilingual inscriptions in the Archaeological Museum of Piraeus, in Greek and Aramaic, testify to the presence of Phoenician traders, who also left more strictly epigraphic traces. (Conversely, Greco-Aramaic stelae in the Archaeological Museums in Istanbul may attest Greek or partially Greek settlements in the Persian empire.) An inscription of the period of Alexander, from the Piraeus, records the response of the Athenian Assembly to the request of some merchants from Cyprus for permission to build a sanctuary to Aphrodite (the goddess, born in the sea, allegedly stepped ashore on Cyprus). The inscription mentions, as a precedent for the request, the Temple of Isis founded by the Egyptian community.

Foreign cults of this kind were not by any means brand-new in the late 5th century; if they seem so, it may be because that period is so much better documented than the early part of the century. But they may have increased in number in Greece as a result of the geographically extensive campaigning of the Peloponnesian War and even the period of the Athenian empire. The cult of Adonis is referred to in Plutarch's *Life of Nicias,* which also mentions the Ammon oracle. Thracian as well as Egyptian cults arrived in Greece in the late 5th century. The cult of the Thracian goddess Bendis at Piraeus features in the first page of Plato's *Republic;* Bendis was perhaps a female counterpart to the Thracian Hero. Cults were both imported and exported: one of the vessels from Rogozen depicts the Greek myth of Heracles and Auge, labeled as such. This is a reminder that the old Olympian cults remained strong. In fact, some of the best evidence for traditional Greek religion comes from this period; it was the century of the highly informative and basically conservative Attic deme calendars (*i.e.,* lists of festivals, chronologically arranged through the year) and the period when inscriptional information about the great Panhellenic sanctuaries entered its richest phase.

Foreign cults

Mercenary service, as well as organized campaigning, must have helped to raise consciousness of such foreign cults as those of Isis or Bendis. Greeks often served in Thrace in the late 5th and the 4th centuries; Xenophon, for example, was there at the beginning of the 4th century and heard the so-called "Ballad of Sitalces" (a 5th-century Thracian ruler who is featured in Thucydides) sung at a banquet in Paphlagonia.

Mercenaries constituted one category of Greeks who strayed away from their cities; they were a potentially disruptive force, whether from the point of view of polis-

minded Greeks or of autocrats like Artaxerxes III or Alexander the Great. Nobody, however, could dispense with them. The Persian kings used Greek mercenaries in their repeated attempts to recover Egypt in the 4th century—but so did the defending Egyptians. How far inside the Persian empire these Greek mercenaries penetrated is an intriguing question. An inscription first published during World War II appeared to attest a group of Greek mercenaries on an island in the Persian Gulf in the period before Alexander, but it is possible that the text is actually early Hellenistic. Even Spartans like Agesilaus near the end of his life and Thebans like the general Pammenes in the 350s had to hire themselves out to Persian paymasters, whether loyalist or insurrectionist. (It would be better to speak, in this context, not of mercenaries but of "citizen-mercenaries" because these Thebans and Spartans did not cease to belong to their home cities.) The military monarchies of Dionysius and Philip were to some extent propped up by mercenary forces, whose loyalty was not subject to political but only to financial blandishments. This leads to the conclusion that the mercenary soldier valued his booty (*aposkeue,* literally "baggage") more than he valued his commander. One of the early successors of Alexander the Great, the Greek Eumenes of Cardia, was in effect traded by his troops to a rival for gain. Already under Alexander the elite troops known as "Silver Shields," or *argyraspides,* had taken their name from the conquered Persian treasure of precious metal.

Not all interchange between poleis, or all emigration from the polis into nonpolis areas of settlement, however, was of the haphazard kind caused by mercenary service or the peripatetic life-style of artists and craftsmen. Rather, the poleis themselves promoted much organized activity.

Colonial connections

First, old ties might be strengthened by renegotiation, or more explicit reaffirmation, of old colonial connections. Inscriptions survive from the 4th century that accord rights of citizenship on a footing of mutuality, for instance, between Miletus and Olbia and between Thera and Cyrene. Some old connections of alliance might be inflated into a pseudo-colonial link. Thus, Hellenistic Plataea, as noted earlier, called itself a "colony" of Athens, which strictly it was not. This claim may well go back to the 4th century, and there is good evidence for other such fabricated claims of kinship in the latter part of this century. An inscription, for example, asserts a colonial connection between Argos and Aspendus in Pamphylia. This is certainly unhistorical but can be explained from the greater prominence enjoyed, in the Hellenistic and Roman periods, by Argos. The reason was that Argos could itself claim a connection with the Macedon of Alexander, and this kind of connection was desirable for obtaining privileges from him or from his successors.

The founding, building, or synoecizing of new cities was another way in which mobility of population was actually encouraged by the poleis themselves. The process is traditionally (and rightly) associated with Alexander the Great himself, but the emphasis is unjust to some innovatory activity in the later 5th and 4th centuries both by individuals (not least Philip) and by cities.

In the late 5th century Olynthus had been synoecized into existence by Perdiccas of Macedon, and the Rhodians had merged the three cities of their island into a new physical and political entity. The same was done in the 360s by the communities of the Dorian island of Cos. Mausolus' new capital of Halicarnassus was the result of a synoecism in which Greeks and native Carians ("Lelegians") were integrated into a new city, which was physically beautified with monumental buildings. Moreover, one can make a case for associating Mausolus with the various refoundations or moving of sites that different kinds of evidence suggest took place at Priene, Erythrae, and Heraclea. Epaminondas' interventions in the Peloponnese led to major urbanization projects at Messene and Megalopolis.

More traditional methods of moving people, such as colonization, were also used; at the beginning of the 4th century Xenophon includes a warm and lyrical description in the *Anabasis* of a site called Kalpe on the Black Sea, praising its situation, fertility, and relative remoteness from rival and established Greek cities in the vicinity. This gives substance to the suspicion that what Xenophon was really trying to do was found a colony of Archaic type—the Euboeans of the 8th century would have jumped at a site with Kalpe's advantages of situation. In the 340s Timoleon of Corinth effected a kind of recolonization of Syracuse from the old mother city; he took with him many refugees and brought prosperity back to an island much battered by internal dissension and endless wars with the Carthaginians—against whom he himself scored some notable successes. Athens sent a colony to the west in the time of Alexander and the corn shortage; it was led with symbolic or sentimental appropriateness by a man called Miltiades (the name of the 6th-century founder and dynast ruling in the Chersonese), who went to the Adriatic region. The Adriatic seems to have been a favourite colonizing focus in this period: the scale and even reality of Dionysius' interventions there are controversial, but an inscription gives evidence of a Greek colony on the island of Black Corcyra. The great colonizing surge of the 4th century came, however, in the wake of Alexander; once again, the Ionian Greeks took the lead. Just as, on Thucydides' evidence, they had colonized Ionia itself even before the organized phase of colonizing activity in the 8th century.

Grants of citizenship

Also in the 4th century a great number of citizenships were granted to individuals from whom favours were expected or by whom they had already been conferred, or both. (One standard motive, occasionally made explicit, for the recording of such honours in permanent form was to induce the recipient to continue his generosity.) Most of the evidence is Athenian, but the phenomenon was surely not confined to Athens. Even Persian satraps like Orontes could be enrolled as Athenian citizens, not to mention Macedonians like Menelaus the Pelagonian, a king of the Lyncestians (an independent Macedonian subkingdom until annexed by Philip). This man received citizenship in the 360s because he was reported by the Athenian general Timotheus as helping Athens in its wars in the north. A further and frequent motive for such honours, and one that anticipates the Hellenistic age, is an expression of gratitude for gifts of grain. The Spartocid kings of the Bosporus (southern Russia) were honoured because they had promised to provide Athens with wheat, as their father Leucon had done before them.

This kind of benefaction is called *euergetism* (the word derives from *euergesia,* or "doing good deeds"). Now that Athens no longer had the naval power to direct all grain forcibly toward its own harbours, much had to be done by exploiting benefactors. *Euergetism* of this sort, however, was not entirely new: as early as 444 BC, Egyptian grain in large quantites had been sent by a rebel pharaoh at a time when Athens was certainly not (as it gradually became) a city armed merely with a cultural past and a begging bowl.

CONCLUSION

No treatment of the main period of Greek civilization should end without emphasizing the continuity both with what went before and with what came after. Continuity is clearest in the sphere of religion, which may be said to have been "embedded" in Greek life. Some of the gods alleged to have been relatively late imports into Greece can in fact be shown to have Mycenaean origins. For instance, one Athenian myth held that Dionysus was a latecomer, having been introduced into Attica from Eleutherae in the 6th century. There is reference to Dionysus (or *di-wo-no-so-jo*), however, on Linear B tablets from the 2nd millennium BC. Looking forward, Dionysus' statue was to be depicted in a grand procession staged in Alexandria in the 3rd century BC by King Ptolemy II Philadelphus. (The iconographic significance of the king's espousal of Dionysus becomes clear in light of the good evidence that in some sense Alexander the Great had identified himself with Dionysus in Carmania.) Nor was classical Dionysus confined to royal exploitation: it has been shown that the festivals of the City Dionysia at Athens and the deme festival of the Rural Dionysia were closely woven into the life of the Athenian empire and the Athenian state. Another Athenian, Euripides, represented Dionysus in a less tame and "official" aspect in the *Bacchae;* this Euripidean

Dionysus

Dionysus has more in common with the liberating Dionysus of Carmania or with the socially disruptive Dionysus whose worship the Romans in 186 BC were to regulate in a famous edict. The longevity and multifaceted character of Dionysus symbolizes the tenacity of the Greek civilization, which Alexander had taken to the banks of the Oxus but which in many respects still carried the marks of its Archaic and even prehistoric origins. (S.H.)

Hellenism

POLITICAL DEVELOPMENTS

Alexander's successors. Nothing shows the personality of Alexander the Great more clearly than the way in which people who had seemed pygmies at his side now became leaders of the world he had left behind. Blood still counted: the only male relative, a mentally impaired, illegitimate son of Philip, was proclaimed king as Philip III Arrhidaeus (*c.* 358–317), together with Rhoxane's son Alexander IV (323–310), born after his father's death in August; both were mere figureheads. For the moment Antipater was confirmed in authority in Macedon and Greece. At Babylon power was shared by two senior officers, Perdiccas (*c.* 365–321) and Craterus (*c.* 370–321). By common consent, Alexander's ongoing plans were abandoned. His generals had to be content with the office of governor. Antigonus Monophthalmos ("The One-eyed"; *c.* 382–301), like Antipater, was not in Babylon at the time of Alexander's death in 323. For almost 10 years he had been governing Phrygia and had shown himself a brave soldier and competent administrator. His firmness and tact were popular with the Greek cities. Of the generals in Babylon, it was Ptolemy (*c.* 367/366–283) who calculated from the first that the empire would not hold together. He secured for himself the governorship of Egypt, where he aspired to set up an independent kingdom. Lysimachus (*c.* 360–281) was given the less attractive assignment of governing Thrace. Two of the others, noted for their physical and military prowess, Leonnatus and Seleucus, waited on events. The soldiers discounted Eumenes of Cardia, who bore the main responsibility for civil administration, but he knew more about the empire than anyone else.

An uprising by Greek mercenaries who had settled in Bactria but wanted to return to Greece was crushed. Trouble in Greece, led by the Athenians and aimed at liberating the cities from Macedonian garrisons, was tougher to control. Sparta refused to participate, as did the islands, but a coalition of Athens with Argos, Sicyon, Elis, and Messenia, supported by Boeotians, Aetolians, and Thessalians, was a formidable challenge to Antipater's authority. For a time Antipater was hard-pressed in Lamia (the war of 323–322 is known as the Lamian War). Leonnatus intervened, nominally in support but in fact ambitious to usurp Antipater's power; he was killed in action, however. In the end Antipater won, Athens capitulated, and Demosthenes (the voice and symbol of anti-Macedonian feeling) committed suicide. Antipater reestablished Macedonian authority autocratically, with no nonsense about a "free" League of Corinth.

The Lamian War

The story of the jockeying for power during the next two decades or so is inordinately complex. First Perdiccas, governing in the name of the two kings with the support of Eumenes, was charged with personal ambition and was assassinated. The armies made Antipater regent (Craterus had been killed in battle), and Antigonus, with Antipater's son Cassander (*c.* 358–297) as second-in-command, was placed in charge of the armies in Asia. Ptolemy was secure in Egypt; Seleucus (*c.* 358–281), governor of Babylon, and Lysimachus in Thrace continued to watch and wait; and Eumenes, a non-Macedonian with a fortune behind him, could claim to represent the kings against the ambitions of generals and governors.

Then, in 319, Antipater died and was succeeded by a senior commander but maladroit politician named Polyperchon, who tried to win the Greeks of the mainland by a new proclamation of their liberties. The result was that the Athenians used their freedom to execute the pro-Macedonians, including the worthy but compromising Phocion. War flared up. Eumenes, allied with Polyperchon, challenged Antigonus and secured Babylon, but he was betrayed and killed in 316. Seleucus escaped to Egypt. Polyperchon's position was weak, and he was soon ousted by the able, up-and-coming Cassander. In becoming master of Macedon and most of Greece, Cassander rebuilt Thebes and put the Aristotelian Demetrius of Phalerum in charge of Athens. Olympias, Alexander the Great's terrible mother, had eliminated Philip III. Cassander had her put to death, while keeping Rhoxane and Alexander IV under his protection—or guard.

Antigonus was now the dominant figure of the old brigade. Cassander, Ptolemy, and Lysimachus formed a coalition against him. For four years (315–311) they fought indecisively. Antigonus showed himself energetic, resourceful, and imaginative, but he could not strike a decisive blow. The only major change came in the brilliant coup by which Seleucus succeeded in recovering Babylon. In 311 the four leaders agreed to divide the world, leaving Ptolemy with Egypt and Cyprus, Antigonus with Asia, Lysimachus with Thrace, and Cassander with Macedonia and Greece, but only until Alexander IV came of age in 305. Seleucus was left out.

Royal blood, however, was quickly forgotten in the pursuit of power. Cassander murdered Rhoxane and young Alexander in 310, soon after Antigonus had vainly tried to crush Seleucus. Seleucus, however, held on to a damaged Babylon and the eastern provinces, except for India, which he had to yield to the Indian king Chandragupta. Antigonus now had the effective support of his brilliant son Demetrius (336–283), known as Poliorcetes, or Besieger, who ousted the other Demetrius and restored the democracy and eventually the League of Corinth; he was hymned with divine honours and given the Parthenon as his palace. Demetrius, also in 306, crushed Ptolemy in a naval battle and secured Cyprus and the Aegean, though he failed in a famous siege of Rhodes (305–304). Antigonus and Demetrius now proclaimed themselves joint kings in succession to Alexander. Antigonus, however, failed to conquer Egypt, and the other rulers also took the title of king. Cassander, Lysimachus, Seleucus, and Ptolemy formed an alliance against Antigonus and Demetrius, and at Ipsus in 301 the allies, with the help of a force of elephants brought from India by Seleucus, defeated and killed Antigonus. Demetrius escaped, retaining Tyre and Sidon and command of the sea. Lysimachus took large portions of Anatolia; Seleucus assumed control over Mesopotamia and Syria, except for a part in the south occupied de facto by Ptolemy; and Cassander was content with Macedonia and parts of Greece.

Murder of Rhoxane and Alexander IV

Cassander, who was a statesman, had founded two great cities, Cassandreia and Thessalonica, as well as rebuilding Thebes. His death in 297 was a prelude to more disturbances. Demetrius conquered most of Greece and secured Macedonia in 294, but he was ousted in 288 by Lysimachus in alliance with King Pyrrhus of Epirus (319–272). Demetrius now concentrated all his forces on winning Asia and all but succeeded. He fell ill, however, and surrendered to Seleucus, who gave him every opportunity to drink himself to death. The stage was set for a confrontation between Lysimachus and Seleucus.

Ptolemy gained command of the sea by Demetrius' fall. He died in his bed, the only one of Alexander's successors to do so, and was succeeded peacefully by his son Ptolemy II Philadelphus (308–246). However, a son by his first wife, Ptolemy Ceraunus, the Thunderbolt (grandson of Antipater), was stirring the waters round Lysimachus, and the latter soon lost support. Seleucus defeated and killed Lysimachus, and Alexander's empire, except for Egypt, seemed to be his for the asking. Lysimachus' army, however, supported Ceraunus, who assassinated Seleucus in 281. Seleucus' son by a Sogdian noblewoman succeeded him as Antiochus I (324–261). In Greece proper the strongest powers were Antigonus Gonatas (*c.* 320–239), son of the brilliant Demetrius and himself a man of high character, ability, and culture, and Pyrrhus, king of Epirus. Pyrrhus was about to embark on his ill-starred expedition to Italy, where he soundly defeated the growing power of Rome but at an enormous cost to himself.

At this point, migrating Celts under the command of

Bolgius and Brennus caused an added complication, not least by the defeat and death of Ceraunus. Brennus pushed down into Greece but was repulsed by the Aetolians. The dangers posed by the invading Celts led, in 279, to a treaty between Antigonus and Antiochus, who agreed not to interfere in one another's spheres of influence. Each won a decisive victory over the Celtic invaders, who eventually settled in Serbia, Thrace, and Galatia in central Anatolia. Antigonus was able to secure Macedonia. Lysimachus' kingdom was never revived. The three centres of power were Macedonia, Syria, and Egypt.

The mid-3rd century. The power of the rulers was not yet secure. Ptolemy II had already launched an offensive after the death of Seleucus and somehow secured Miletus. He made a new drive in 276 to gain Seleucid Syria only to be repulsed. About that same time, however, he renounced his first wife and married his sister Arsinoe, who was actually widow to both Lysimachus and Ceraunus. She was a woman of dynamic authority who inspired Ptolemy's armies to sweep up the coast and secure Phoenicia and much of coastal Anatolia. Her brief years were years of brilliant culture. When she died on July 9, 270, the court poet Callimachus wrote a poem on her deification.

Arsinoe

In the west, Pyrrhus, returning to Epirus full of thwarted ambition, overran Macedon but abandoned it to attack southern Greece. He failed, however, to take Sparta and died in street fighting in Argos, after being struck to the ground by a tile hurled down by a woman watching from the roof. Pyrrhus had fostered the Hellenization of northwestern Greece and built the magnificent theatre at Dodona; he was more than a military adventurer.

Antigonus was influenced by stoic philosophy (see below); he had a high sense of duty and once said that the power of kings was merely a spectacular form of servitude. He also was a friend of the poet Aratus. There was no serious challenge to his power in the north. In the south, Athens, led by the handsome Chremonides, allied with Sparta and other cities against him; the alliance was backed by Egypt and received some support from Epirus. The war was hard-fought for four years (266–262), but the alliance fell apart. The political power of Athens was finally broken, but the city survived as a cultural centre. Antigonus left Sparta to itself and placed dictators (tyrants) of his own choice in other cities.

Antiochus I of Syria died in 261. He was succeeded by his son Antiochus II (287–246), who formed an alliance with Antigonus against Ptolemy II. In the Second Syrian War (259–255), Antiochus recovered most of the coast of Anatolia and Phoenicia, while Antigonus won a naval victory and with it command of the sea; he even was able to put a half-brother into power in Cyrene. The death of Antiochus II in 246, however, brought on a fresh power struggle in Syria, and Ptolemy III Euergetes (*c.* 284–221), succeeding his father in the following year, was able to march through the distraught realm. Seleucus II Callinicus (*c.* 265–225) eventually restored stability and recovered some but not all of the lost territory. Yet he was again challenged by civil war and had to abandon Bactria, Parthia, and the eastern provinces (Cappadocia had already been lost before the civil war).

The weakness of the Seleucids brought a new power onto the scene. Pergamum had great resources in silver, agriculture, and stock breeding but had not come to marked prominence. Attalus I Soter (269–197), who ruled from 241 to 197, made Pergamum a great power. He defeated the resurgent Celts of Galatia, took the title of king, for a period held mastery of much of Anatolia, intervened in the west, and all the while made his city a major centre for literature, philosophy, and the arts.

During the middle of the century some remarkable developments in confederation occurred on mainland Greece. Epirus had been a form of confederacy between Molossians, Thesprotians, and Chaonians. Pyrrhus had established an autocratic monarchy, but after his death in the 230s the people reverted to a federal constitution. In Boeotia, a confederacy composed of officials predominantly from Thebes (the largest city in a system that gave all citizens the right to vote in the primary assembly) modified its pattern to grant equality to the constituent cities regardless of size. In Aetolia, there was a confederacy with a strong primary assembly that met twice a year and a council with proportional representation of the member states based on each state's military contingent; the existence of tribal districts intermediate between the cities and the whole confederacy was an unusual feature. Neighbouring Acarnania also had a federal constitution. The two neighbours were generally hostile, but at one point they actually agreed on limited mutual rights of citizenship.

The best-known of the confederacies was the Achaean League. It had existed earlier, to be revived in 280 by the cities of Dyme, Patrae, Tritaea, and Pherae; it was joined by Aegium, Bura, and Cerynea. "For the first 25 years," wrote the historian Polybius, "the above-mentioned cities shared in a confederacy, appointed a common secretary according to a rota, and two generals. After that they took a fresh decision to appoint a single general and to entrust him with plenary authority. Margus of Cerynea was the first." There were also 10 magistrates called *demiourgoi*. Then, in 251, the Greek statesman Aratus (271–213), incorruptible, adventurous, persuasive, skilled in diplomacy, passionately attached to freedom and implacably ambitious for his own position, rid his native Sicyon of its tyrant and brought it into the league. By 245 he was elected general and held the office in alternate years. Aratus heartily loathed tyrants and Macedon alike. A notable guerrilla fighter, he led the league in the work of liberation, freeing Corinth and winning Megara and some cities of the Argolid but not Argos or Athens. Then he clashed with the revolutionary nationalism of Cleomenes III of Sparta (*c.* 260–219; see below), and, rather than seeing his life's work imperiled by Cleomenes' revolution, he preferred to sell it back to the imperialists of Macedon. Macedon came and conquered. Aratus and the league were allowed to retain a shadow of independence, but no more than that. The league, however, remained intact. Executive power lay with the Council, which seems to have been a large body constituting a kind of representative government. What the Achaean League did, for a limited period over a limited area, was to combine the distinctive character of the city-state with a wider vision. On the coins the local Aphrodite of Corinth and Hera of Argos yield place to the more widely recognized Zeus Homagyrius and Demeter Panachaea. According to Polybius, the whole Peloponnese during the most important phase of the Achaean League could be considered a single polis.

Aratus

Sparta, always different from the rest of Greece, was a shadow of its former self. There were no more than 700 Spartan citizens, and the land, far from being equally distributed, was in the hands of only a few. Agis IV, coming to power in 244, essayed economic and social reform by abolishing debts and redistributing land. He succeeded in the former but was killed by those whose power he threatened. His widow was married to Cleomenes, son of the other king, Leonidas II. She, however, won him to the need for revolution. In this she was supported by Cleomenes' stoic tutor Sphaerus, who seems to have read a remarkable utopian narrative composed *c.* 250 by an otherwise obscure author named Iambulus. Cleomenes came to the throne in 235; in 227 he began to break the power of the oligarchy within the aristocracy, abolish the debts owed by poor farmers to rich landlords, and redistribute the land. He also reintroduced the common meals and restored the simplicity of life and the education for character that were traditional in Sparta. Cleomenes III combined a narrow Spartan nationalism with a visionary idealism. The revolution spread; everywhere there was demand for "division of land and cancellation of debts." Cleomenes, however, was stopped by Aratus, an adamant opponent of his reforms, the Macedonians were called in, and at Sellasia, in the summer of 222, the Spartans were beaten and Cleomenes forced into exile, where he died.

The coming of Rome (225–133). In the 3rd century, Rome had been encroaching on the Greek settlements of southern Italy and Sicily. Pyrrhus, as noted above, had been called in by Tarentum in the Tarentines' fear of Rome. Hieron (*c.* 306–215), a Syracusan supporter of Pyrrhus, seized power in his city; he was made king in 269 and actually reigned for 54 years. For a year or two he

continued to oppose Rome, but then he formed an alliance with it, helping it in its wars with Carthage. Farther away yet, Massalia (modern Marseille), an outpost of Greek culture, took care to maintain good relations with Rome; at the same time, it maintained a strong independent navy and a stable oligarchic government. (Massalia is a classic example, often forgotten, of the durability of the Greek city-state in the Hellenistic age; even in 121 BC, when the Roman province of Gallia Narbonensis was established, Massalia was still an equal ally of the Roman Republic.)

In the late 220s new monarchs acceded to the throne in the three great kingdoms of Syria, Egypt, and Macedon, and Polybius chose that point for the formal start of his history. Antiochus III (*c.* 242–187), called the Great, succeeded his brother Seleucus II in Syria, and from the first he showed a desire for imperialist expansion. His attempt to conquer Egyptian territory in the Palestinian area in the Fourth Syrian War (219–216) was foiled at the battle of Raphia. His campaigns in the east were more successful: he secured Armenia, Parthia and Bactria became his vassals, and he carried out impressive demonstrations near the northwestern frontier of India and across the Persian Gulf. He turned to adventures in Europe but came up against a Rome resurgent after its war with Hannibal; by the peace of Apamea in 188 he was confined to his still considerable Asian domains. In Egypt, Ptolemy IV Philopator (*c.* 244–205) succeeded to power in 221. He repelled Antiochus III at Raphia with Egyptian soldiers, and his reign was marked by the power of native Egyptians and of Nubian rulers in the south. He died in 205, leaving a five-year-old son. There occurred an uprising, which deposed his minister Agathocles, and disturbances throughout the reign.

Philip V of Macedon

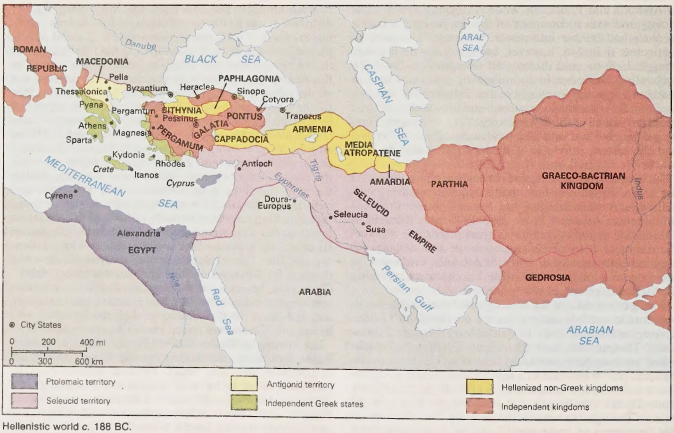
Philip V of Macedon and quite capable on the battlefield, he showed unsound judgment and lacked stability of temperament. Like Antiochus, he had expansionist ambitions, but he supported Hannibal against Rome and was roundly defeated by the Romans at Cynoscephalae in 197.

Rome was almost forced into the Greek world. In 229–228 and again in 219 it had been campaigning against pirates in Illyria. Then, from 218 to 201, it was preoccupied with and became drained by the Second Punic War with Hannibal. Even so, Rome kept Philip V at bay and, once Hannibal was eliminated, defeated him in the Second Macedonian War. Rhodes and Pergamum had checked Philip's enterprises in the Aegean but were understandably nervous about his future intentions. They called in the Romans, who were equally suspicious of Philip. Their victory over him at Cynoscephalae, where the Macedonian phalanx of heavy infantry showed that it was hard to beat if it kept its ranks but vulnerable if it did not, demonstrated Rome's supremacy. Rome, however, annexed no territory; the narrow oligarchy governing Rome had no desire to take on administrative responsibilities that might require extending the circle of those in power. The young commander Titus Quinctius Flamininus (consul in 198) was a philhellene. At the Isthmian Games in 196 he proclaimed the freedom of Greece. A priesthood to him was set up at Chalcis, which still survived in Plutarch's time, and a paean was composed to Titus, Zeus, and Roma, ending "Hail Paean Apollo, hail Titus our Saviour" (or "Liberator"). He checked the ambitions of Nabis of Sparta, who combined the revolutionary program of Cleomenes III with imperialism and cruelty. Yet in 194 all Roman troops were withdrawn from Greece.

The next challenge came from Antiochus, as already indicated. The Romans returned to Greece to fight him. They defeated him in Asia, strengthening Pergamum and Rhodes at his expense but annexing no territory themselves. Then Perseus (*c.* 212–165), son to Philip V, succeeded to his throne and power in 179. He secured his position by dynastic marriages; he wedded the daughter of Seleucus IV (*c.* 218–175) and was allied by marriage to Prusias I Cholus of Bithynia. In addition, he used diplomacy to extend his influence. Nevertheless, in 172 Eumenes II of Pergamum (d. 159), who had succeeded his long-lived father in 197 and who was a great builder in his capital, felt threatened by the growth of Macedonian power and appealed to Rome. The result was the so-called Third Macedonian War (172–168), which ended with the defeat of Perseus by Lucius Aemilius Paullus at Pydna. Macedonia was divided into four republics—and yet again the Romans withdrew without annexations. If Rome, as its enemies avowed, was a dragon, it was a reluctant dragon.

Meantime, Antiochus IV Epiphanes of Syria (*c.* 215–163) had come to power in 175. He had been a hostage in Rome and was a passionate philhellene; he paid lip service to the political traditions of both Athens and Rome. The Romans, however, prevented him from annexing Egypt and Cyprus, which he had invaded in 168.

Antiochus actively pursued a policy of Hellenization as a means to unify his kingdom. This policy, however, led to an uprising in Judaea, though it should be emphasized



Hellenistic world *c.* 188 BC.

that it was a pro-Syrian party among the Jews that applied to the king for permission to build a gymnasium, with all that this implied. Party conflict among the Jews—*i.e.*, the supporters of Hellenization and the orthodox Jews who fiercely opposed it—was a major factor in the disturbance. Equally, Antiochus' sense of his own divinity, represented by the title Epiphanes (God Manifest), was unacceptable to the orthodox Jews who recognized the absolute claims of the God of Israel. Antiochus forbade the practices of the Jewish faith and placed an altar to Olympian Zeus ("an abomination of desolation") on the altar of the temple. Resistance flared up, first passive, then, under the leadership of Judas Maccabaeus (who made "a league of amity and confederacy" with the Romans), active and military. The details of the conflict as it spread over decades and the reigns of successive rulers of Syria are complex: suffice it to say here that for virtually a century the Jewish people enjoyed a large measure of de facto independence.

Antiochus and Judas Maccabaeus

By 146 the Romans were impatient with Greek instability, and at the same time they were determined to have done with Carthage. The city was razed and a province established in the fertile farmland of modern Tunisia. A pretender, who had arisen in Macedon, invaded Thessaly; he was defeated, captured, and executed, and Macedonia was annexed as a Roman province. The Greeks clashed with the Romans; patriotic sentiment ran high but to no effect. The Romans treated Corinth as Alexander had treated Thebes—they leveled it. In the rest of Greece the leagues were dissolved, democracies abolished, and power placed with the rich. Intercity peace was established and left to the governor of Macedonia to enforce. The ironic result was that the city-states had, imposed from outside, a degree of autonomy and peace they had previously lacked. Then, in 133, Attalus III of Pergamum (*c.* 170–133) bequeathed his kingdom to Rome—an odd, though perhaps realistic bequest. It aroused opposition, led by a pretender named Aristonicus, who was driven by a combination of personal ambition, nationalist resentment, and utopian idealism. The movement was backed by a stoic philosopher named Blossius, who had been concerned with the reforms of the Gracchi in Rome. It spread among the oppressed and aimed to establish a utopian "City of the Sun." Roman military power, however, was too strong. Aristonicus was defeated and killed; Pergamene territory became the Roman province of Asia.

For the most part the story of the kingdoms of Egypt and Syria during the 2nd and 1st centuries was one of stormy and deeply divisive feuds. In Egypt brother-and-sister marriage in the royal house was frequently practiced. The rulers were for the most part an undistinguished lot, yet the country remained wealthy, and there was expansion to the south. In Syria civil war and division seemed to be the rule rather than the exception. Antiochus VII Sidetes (*c.* 159–129), after a victorious campaign in Mesopotamia, Babylonia, and even Media, looked briefly as if he might restore the lost glories. The Parthians, however, rallied, surprising and killing him in the winter of 130–129, and regained all he had recovered. Thereafter the kingdom became weak and divided, and neighbouring states were constantly gnawing at its edges. Far to the east the Greek dynasty that had ruled Bactria since about 256 was coming to an end by the middle of the 1st century. In western India, however, Menander, a hero of Indian legend, was in power; the art of the Gandhara region (present northwestern Pakistan) shows marked Greek influence.

Mithradates (Mithridates) VI Eupator of Pontus (*c.* 132–63 BC) was still a minor in 120—the year that his father was murdered—when he was named joint ruler with his mother and brother. For some years he was a refugee from his mother's power. Then, in a sudden sally, he secured the throne, imprisoned his mother, killed his brother, and married his sister. Pontus, sprawling along the southern coast of the Black Sea, included Greek colonies and a native population; the largest section of the people, including the rulers, were Iranian. Mithradates was able, cunning, and ambitious. He secured money and men by expanding to the north and then turned to Anatolia, the Aegean islands, and even Greece, where the financial oppression of the Romans made him appear a liberator. The Romans defeated him time and again, but he showed a subtle resilience until his final defeat by Gnaeus Pompeius Magnus (106–48 BC). In 67 Pompey made his greatest contribution to peaceful trade and development by his systematic destruction of the pirates. He put an end to the danger from Mithradates, who was driven from his kingdom and committed suicide in 63. Pompey in his celebrated settlement of the East annexed Syria as a Roman province, settled Judaea, and planted Roman colonies.

Henceforth the Greek world was dominated by Rome. Julius Caesar and Pompey faced one another at Pharsalus in Thessaly in 48. Mark Antony and Octavian faced Marcus Junius Brutus and Gaius Cassius Longinus at Philippi in Thrace. The brilliant Cleopatra VII (69–30 BC), last of the Greek Ptolemaic dynasty in Egypt, was ambitious to rule the world. In the realism of power politics she had to conquer Rome: the path lay through marriage with whoever held the power there. The surviving portraits show that she was no great beauty. Nonetheless, she charmed Caesar and held Antony in the power of her personality. Yet she backed the wrong man. A third conflict for the mastery of the world in two decades was held in Greece, culminating in the naval battle of Actium off the western coast in 31. The victor was Octavian (63 BC–AD 14), the future Caesar Augustus. The last kingdom of Alexander's successors fell to Rome.

Cleopatra

The Greek world under the Roman Empire. Under Augustus, Macedonia, though including Thessaly, was separated from a new province of Achaea with its administrative centre in Corinth; both provinces were assigned to the Roman Senate. Thrace remained a kingdom and was not annexed until AD 46, when it became a province under an imperial procurator. Asia was a province, incorporating the western coast of Anatolia and reaching into the interior. Bithynia-Pontus stretched along the northern coastline. There was a third area of special command in Cilicia, but this did not last; part of it went to Syria, part to a new province, Galatia (25 BC), and part to small vassal states. The imperial province of Cilicia dates from AD 72. Lycia et Pamphylia became a separate province under Claudius in AD 43; and Cappadocia had been annexed earlier under Tiberius in AD 17. Cyprus constituted a province, at first under the emperor, but later it was transferred to the senate. Crete and Cyrene formed a single province. Syria was the most important of the eastern provinces. Finally there was Egypt, an imperial preserve and vital for the grain supply and revenue of the empire.

During the next century and a half, four major factors affected the eastern half of the empire. First, a whole series of earthquakes and other calamities devastated the cities of Anatolia. (Strabo, the aforementioned Greek historian and geographer, has an appalling account of the eruptions around Philadelphia, which drove the inhabitants into the open country as refugees.) The Roman emperor Tiberius (ruled AD 14–37) met these disasters with constructive aid, and 12 cities set up a record of his benefactions, calling him "the simultaneous founder of 12 cities." Second, there began to be trouble with the Jews, not just in Judaea but in Alexandria and elsewhere as well. The Roman emperor Caligula's demands for worship led to an embassy to Rome, recorded by the great Hellenized Jewish scholar Philo Judaeus (*c.* 30 BC–AD 45). The emperor's death resolved the problem, but even the more tolerant Claudius (ruled AD 41–54) had to intervene in Alexandria. The wars of 66–70 and 132–135, revolts against Roman rule in Judea, had the effect of further dispersing the Jewish people around the empire. Third, Nero (ruled 54–68), patron of the arts and philhellene, made a triumphal tour of Greece, dancing, singing, competing, and carrying away the prizes. The four great festivals at Olympia, Delphi, Nemea, and the Isthmus were crowded into one year for his sake. "The Greeks are the only people who understand how to be an audience," he said. He proposed and began a Corinth canal and proclaimed the freedom of the Greeks, which, in effect, meant reduced taxation. Finally, there was the eastern frontier and the power of the Parthians, and subsequently Sāsānian Persia. Armenia was a focal point. The rulers in general walked a tightrope between the great powers with skill. Trajan (ruled 98–117), however, fol-

lowed a strong-arm policy. He annexed Armenia, making it a province, did the same to Mesopotamia and to Adiabene, and captured Ctesiphon. On his coins he put the inscription *Parthia capta* ("Parthia conquered"), followed by *rex Parthis datus* ("the king given to the Parthians") as he imposed a puppet monarch. He dreamed of being a second Alexander, but he died, and Hadrian (ruled 117–138) gave up the three new provinces, retaining only a fourth, Arabia.

A period of peace was then followed by one of chaos. During the reign of Marcus Aurelius (ruled 161–180), Ctesiphon was again taken; but there also was a disastrous plague spread through the Greek world and even to Italy and Rome. The 19th-century classical historian Barthold Georg Niebuhr said that the ancient world never recovered from the blow. In addition, mainland Greece suffered from an incursion of a people called the Costoboci, who even succeeded in sacking Eleusis in 170–171. The accession of Septimius Severus (ruled 193–211), who came from Africa, brought a remarkable coterie of women from Syria to power in Rome. Julia Domna, "Julia the philosopher," was the emperor's second wife. She established a highly cultured court, inviting to it scholars and writers such as Galen and Philostratus from the Greek east. Her sister Julia Maesa and her nieces Julia Soaemias and Julia Mamaea were responsible for the coming to power first of the fantastic Elagabalus and then of the young Severus Alexander.

"Julia the philosopher"

In 224 the Sāsānian dynasty came to power in Persia with an autocratic centralized government upheld by a strong religious commitment. Its rulers aimed to drive the Romans out of Asia; in 256 they ravaged Antioch, and in 260 they captured the emperor Valerian. At Palmyra, an outpost of Greek culture, the remarkable Septimia Zenobia came to power and at one time conquered Syria, Egypt, and much of Anatolia. In 267 the Germanic Heruli actually sacked Athens, Corinth, Argos, and Sparta. But the Romans were resilient. Aurelian recovered the lost ground. He had, however, felt the power of the east, and in 274 he introduced the Unconquered Sun as the supreme god of the Romans. It was clear to the emperor Diocletian that the administrative system had to be changed; he placed two rulers in charge of the east (himself being one of them), and two of the west, with 13 regions and 116 provinces. Nicomedia in Bithynia was chosen as the eastern capital. Constantine (*c.* 285–337), after winning sole power, went further and moved the capital of the whole empire into the east. He thought first of Troy, as Julius Caesar had before him, but in the end he chose Byzantium with its magnificent site where the Bosporus, the Golden Horn, and the Propontis meet. On May 11, 330, he inaugurated New Rome, or Constantinople, to be the capital of the new Christian empire. It was, in a sense, the triumph of Hellenism and ensured the survival of the Roman dominion in the east for more than 1,000 years.

HELLENISTIC CIVILIZATION

Institutions and administrative developments. *The great cities.* The greatest of Alexander's foundations became the greatest city of the Hellenistic world, Alexandria-by-Egypt. It was laid out in the typical Hellenistic grid pattern along a narrow strip between Lake Mareotis and the sea. Canopic Way ran the length of the city, finely paved and nearly 100 feet (30 metres) wide, with seven or more main roads parallel to it. Across it was the shorter Transverse Street, with at least 10 parallel major roads. The city was divided into five regions, known as Alpha, Beta (the Palace area), Gamma, Delta (the Jewish quarter), and Epsilon. The great buildings included the palace, Alexander's tomb, the temple of the Muses, the academy and library, the zoological gardens, the temple of Sarapis, the superb gymnasium, stadium, and racecourse, the theatre, and an artificial mound, the shrine of Pan, ascended by a spiral road. There were two harbours. The famous lighthouse lay on an offshore island. A canal to the Nile helped secure the water supply; there also were rainwater cisterns. The city wall was some 10 miles (16 kilometres) long. It was a cosmopolitan city. The so-called Potter's Oracle described the city as "a universal nurse, a city in which every human race has settled," and Strabo called it "a universal reservoir."

The great Seleucid capital Antioch on the Orontes stood safely some 11 miles from the sea on a major trade route. Originally small, the grid plan with blocks roughly 390 feet by 115 feet was laid out from the first for the expansion over the plain, which eventually took place. A colonnaded street, in Roman times more than 88 feet in width (about one-third carriageway and one-third for each sidewalk), ran the city's length. Aqueducts brought water from the mountains to flow in water conduits along the east-west streets and through terracotta pipes along the cross streets. Like Alexandria, the city was cosmopolitan, and Tacitus speaks of intermarriage between the ethnic groups. When Julia Domna held court, students came from Phoenicia, Palestine, Egypt, Cyprus, Arabia, Cilicia, and Cappadocia. It was a city noted for its luxurious living, as the magnificent mosaics of the Roman period from Antioch itself and the fashionable suburb of Daphne demonstrate. Antioch suffered severely from earthquakes and flooding; thus there was much rebuilding. The population was perhaps 500,000 at its largest.

Antioch on the Orontes

It seems that Antioch was smaller than Seleuceia on the Tigris, the largest of all the Seleucid foundations, with 600,000 inhabitants according to Strabo. Little, however, is known about it, except that it was on a grid plan and had a stone wall on foundations of baked mudbrick. In 143 BC it became an autonomous Greek city under Parthian control.

Pergamum, a small town before it became the capital of the Attalid dynasty, remains one of the most spectacular of ancient sites. The southern face of the acropolis was brilliantly terraced and carried two agoras, stoas, a gymnasium, palaestras, an odeum, temples, and other buildings. Near the top stood the great altar of Zeus with its mighty frieze (now in Berlin) of the battle of gods and giants, and the throne of Satan. The main street leads through a gate to the upper city. There one finds an imposing sanctuary of Athena, the famous library, and a temple of Trajan, on which excellent restoration work has been done. Below is a vertiginous theatre seating 10,000, with a removable stage building on a terrace leading to the Ionic temple, presumably of Dionysus. Much of the remainder of the upper city was occupied by the palace buildings and storehouses. Less can be discerned of the lower city, except for the sanctuary and hospital of Asclepius, founded about 400 BC but developed in the Hellenistic and Roman periods.

The most evocative remains of all the ancient cities are those of Ephesus. It was moved to its longest-lasting site about 290 BC by Lysimachus and built mostly on a grid plan with a wall of more than 5.6 miles. Much of what is visible today dates from the Roman period. By the harbour, today far inland, are the great baths; a broad colonnaded street leads to the theatre, the scene of the silversmiths' riotous protest against the apostle Paul. A cross street passes in front of the theatre, with a huge agora to the south and the imposing Library of Celsus, dedicated in AD 110. From there the slightly eccentric Curetes Street runs eastward. On one side are wealthy houses with mosaics and frescoes, on the other the Baths of Scholasticia and the Temple of Hadrian. Further up the street is the colossal terrace that sustained a Temple of Domitian and leads on to the area of the State agora, the political, administrative, and religious centres, and a magnificent gymnasium. The great Temple of Artemis, a little way off, was one of the Seven Wonders of the World.

The administration of Ptolemaic Egypt. Ptolemaic Egypt represented, in the words of the 20th-century historian Frank William Walbank, "a large-scale experiment in bureaucratic centralism and in mercantilism." There was a constant need to import material not readily available at home, such as the timber and pitch required for warships and the mercantile fleet and also gold. Demetrius, chief executive of the mint in Alexandria, wrote to Ptolemy II's finance minister, Apollonius, in 258 BC about the need to import as much gold as possible. The Ptolemies had a closed monetary system, which required all foreign traders on arrival to change their money. Exports included linen, papyrus, faience, and eventually glass (with

The Ptolemaic economy

a stringent quality control), and, when appropriate, grain. The administrators divided the country into more than 30 regions, or nomes, with smaller divisions into districts and villages. There was military government alongside a complex financial administration responsible for collecting rents and taxes. At the same time, the local finance offices were instructed (a document survives) to encourage the peasants, protect them from disaster, and ensure the sowing of the correct crops. The king, in theory, claimed all the land and let it to peasants on short leases, providing the seed-corn but requiring its equivalent to be returned. The oil-producing crops were state monopolies; so were mines, quarries, salt, nitre, and alum. Other areas of agriculture were controlled by license, taxation, and price-fixing. A surviving letter from a finance minister says, "No one has the right to do what he wants, but all is regulated for the best." Perhaps it was not always as systematic, efficient, and incorrupt as it sounds or as some admirers have proposed. Nor was it all so new. The major change was the imposition of a Macedonian and Greek ruling class, who filled the upper ranks of the civil service. Egyptians held some of the lower posts, but only in the priesthoods could they retain wealth and influence. There was friction at times; for example, a camel driver complained of nonpayment because "I do not know how to behave like a Greek."

Still, there were few slaves outside the cities, and double names attest the gradual acceptance of some Egyptians into the upper echelons of society. The remarkable Cleopatra VII, however, was the first sovereign to learn the native language. When all is said about defects in the administration, Egypt was, and remained, inordinately wealthy, and the Romans were delighted to secure its revenues and its grain.

Military developments. The victories of Philip II and Alexander the Great were made possible by imaginative generalship and inspirational leadership combined with the use of elite troops that were specially trained and equipped. The Macedonian phalanx depended upon a long, heavy spear called a *sarissa.* The troops were organized in battalions of about 1,500 men forming 15 rows in depth. The 11 rows at the rear held their spears vertically, causing them to tower formidably above them. The four front rows held their spears horizontally so that all projected in front of the phalanx. For protective armour they wore helmets, leather corselets, and metal greaves, and each carried a small round shield. The phalanx was virtually impregnable to a frontal attack but could not easily swerve or reverse. The heavy cavalry of the Companions carried a shorter spear and scimitar and wore metal helmets and breastplates. They advanced in the form of a spearpoint, or triangle, so as to break up the opposing line of battle. On the wings of the phalanx were fairly mobile troops: light cavalry, slingers and archers and javelin men, and light infantry.

The successors recruited large armies of 60,000 or even 100,000 men, including many mercenaries. By about 200 BC troops from Greece, Crete, and the Balkans had decreased in number and many more were recruited from the Syrian territories. The mercenaries were not normally trained for the phalanx but were supplementary to it. The employment of mercenaries increased the number of desertions and the amount of looting; this in turn led to the need for more stringent discipline in the field. At the same time, the armies were relatively free from the hatreds liable to arise between highly politicized national forces. Surrender on easy terms followed by ransom tended to be the order of the day.

The armies

Alexander was a great master of siegecraft. He used saps and mines, timbered galleries, catapults and stone throwers, siege towers, scaling ladders, and covers for such operations as filling up ditches or bringing battering rams to bear. These new devices were countered by better walls, towers, ditches, and outworks so that in general the besiegers had to rely on treason, bribery, stratagem, or on starving out the besieged town. Demetrius and Philip V were the only two of the successors who gained much reputation in siege work.

The fleets of the Hellenistic age were smaller in number of boats than those of the Classical period, but the battleships were larger. Ptolemy II's fleet of 336 was smaller than that of Athens in its war with Sparta. The quinquereme, however, was now the standard battleship, and its crew was about double that of the trireme. Even larger vessels were used, such as a 16-oarer with two banks of oars and eight men to an oar. The Macedonian king Antigonus Gonatas had a flagship of the 18-oar type. One even hears of a 40-oarer. In general, the Macedonian navy dominated the Aegean and the Egyptians the rest of the eastern Mediterranean. There were, however, many fluctuations, and Rhodes was never negligible.

Civic structures. Wherever Hellenization was strong, there tended to be support for the institution of the city-state as well as a measure of synoecism, or gathering of smaller communities in a new polis. The Alexandrias were followed by countless towns, to which were given names such as Antiocheia, Seleucias, Laodicea, Ptolemais, Demetrias, or Cassandreia. Some townships that were not essentially Greek, such as Tyre, Sidon, Byblos, Aradus, and Sardis, were nonetheless treated as cities, except for the towns of Mesopotamia. Non-Greek immigrants into Greek cities might be granted their own administrative system rather than being absorbed into the general citizenship; for example, the Jews in Alexandria had their own ethnarch and Council of Elders.

Some of the successors were hostile to the Greeks, notably Antipater and Cassander. All were liable to impose conscription and taxation, though occasionally immunity was granted. The kings exercized control through a resident representative (*epistates*) in the cities, though this was generally handled delicately and diplomatically. Sometimes, however, they preferred to support a puppet dictator. The rights of minting coinage were severely restricted. The apparatus of civic government, however, remained, and, under the Seleucids, decrees were passed by council and assembly in city after city. During the periods of relative freedom in mainland Greece, there was sometimes democracy, and the Ptolemies maintained democracy in Cos. Yet the kings generally, and the Romans after them, encouraged autocratic or oligarchic government. Most cities in mainland Greece and some others, such as Rhodes, Cyzicus, and Byzantium, retained rights of foreign policy, including military action. They also acted to maintain the grain supply, sometimes by the public purchase of grain and its cheap sale or free distribution. The same freedom made possible the remarkable developments in federal government already noted. This in turn led to a great increase in the use of arbitration in the settlement of disputes, which was obligatory within the confederacies or among those cities directly dependent on the monarch and not infrequent outside.

The encouragement by the overlord, whether Greek or Roman, meant changes in the political patterns. These can be seen reflected in Roman times in the works of Plutarch (who, however, idealizes the past to such an extent that one cannot be sure of him as a contemporary witness) or of the Greek rhetorician and philosopher Dion Chrysostom. Plutarch preferred monarchy and was opposed to extending the franchise to all the free population; interestingly, though, he favoured some kind of party system, so that there is more than one policy to choose from. The changes meant a more or less settled ruling class in the cities. There was now no room for demagogy because there was no deme which it made any difference to court. Where the politically ambitious had scope was in deputations to the kings or, later, the Roman emperors. Nonetheless, the path of the ruling class was not always strewn with roses. Its members were expected to bear the brunt of public expenditure, which in the harsher times of the later empire could become burdensome. In questions addressed to an oracle, found at Oxyrhynchus and dating from the late 3rd century AD, the inquiries "Am I to become ambassador?" or "Am I to become a senator?" are not very different from the question "Am I to become bankrupt"? They were dictated by fear, not ambition. Similarly, there are some amusing records of council meetings which show nominees eager to wriggle out of an office that might become expensive, while the

Changes in the political patterns

others blocked their paths of escape and applauded the patriotism of acceptance.

Slavery was virtually universal but varied in its incidence. On the whole, though there were numerous exceptions, Greeks did not enslave Greeks; their slaves came predominantly from Anatolia and Syria, Thrace, the Danube basin, and southern Russia. The main sources were war and piracy, fostered by the work of the slave-dealers. The great centres of the trade were Byzantium and Ephesus, but, from the middle of the 2nd century BC to the middle of the 1st, Delos became the dominant market. In the Greek east, slaves were numerous in the cities; it should, however, be noted that they could hold relatively responsible jobs. There were comparatively few slaves in the countryside. Under the early Roman Empire the supply dwindled with the control of piracy and a long period of peace. Liberal legislation by Claudius in the 1st century AD and by Trajan, Hadrian, and Antoninus in the 2nd gave increasing protection and rights to slaves. The price of slaves rose, which implies that often they could be afforded only for skilled work. In the 3rd century, with frontier wars and brigandage resurgent, the prices dropped somewhat, but demand still outstripped supply. The breeding of slaves continued, and the sale of newborn babies was legalized and that of older children, though illegal, was widespread.

Economic developments. Alexander's conquests had four major effects on the economy. In the first instance, it released a large quantity of silver and gold from the treasuries of Persia. The immediate result was a sharp rise in prices, but, as the surplus funds were absorbed into capital, prices began to fall. Second, the integration of quarreling city-states into a single empire removed some of the obstructions to mutual trade. Third, Philip had already adopted the Attic standard for gold, and Alexander adopted it for silver as well. The successors in general followed, though the Ptolemies preferred the Phoenician standard. The complex needs of money-changing were thus greatly reduced. These two standards held good until some time in the 1st century BC, when the Roman challenge to them triumphed. Finally, and most obviously, the extension of empire meant an extension of trade routes; China became open to trade for the first time and East Africa, Arabia, and India became more easily accessible than before.

Egyptian trade

The Egyptian trade was mainly by sea, featuring the port of Berenice on the Red Sea, while Alexandria was established as one of the greatest mercantile centres on the Mediterranean. Toward the end of the 2nd century BC an Indian at Alexandria explained to Ptolemy VII the secret of the monsoon, which greatly facilitated the sea passage to India and enhanced the importance of Coptos on the upper Nile. The Egyptians also had an eye to the land routes. This explains their desire to command the Phoenician ports, which were not only the terminus of one land route but also producers of woven stuffs and fine dyes.

The keypoint for Seleucid trade was Seleucia on the Tigris. In one direction, the route led to Antioch on the Orontes with branches to Ephesus and Damascus. In the other, there were three routes to India, two by land and one by sea. Alexander's foundation of Alexandria in Areia was important to the trade. Dura Europus on the Euphrates was a fort protecting the lines of trade; it was retained by the Romans. The caravan cities, such as Petra and Palmyra (formerly Tadmor), flourished on the trade. The advance of Chinese military power from Turkistan in the 2nd century BC fostered the trade with China along the famous Silk Road through central Asia. The Chinese exported silk and other textiles, bamboo, and iron and imported vines and other trees and plants, as well as wine, olives, woolen goods, and artwork (which affected Chinese artistic style). The demand for luxury goods in the prosperous days of the early Roman Empire increased the trade with China, India, and Arabia, and an embassy from Marcus Aurelius actually reached China by way of Annam.

Early in the Hellenistic age, the Greek navigator, geographer, and astronomer Pytheas of Massalia embarked on one of the most remarkable feats of exploration. Evading the Phoenician outposts, he slipped through the Strait of Gibraltar, sailed north along the coasts of the Iberian peninsula and France, crossed over to Cornwall, continued around the north of Britain and on to Helgoland, and then returned. The Phoenicians, however, allowed no other ship to pass Gibraltar and the only tangible result of Pytheas' voyage was an increase in the trade in Cornish tin by overland routes through France.

In general the Romans made transport, whether by land or sea, safer and swifter. The Greek Epictetus could say, "Caesar has procured us a profound peace. There are no wars, no battles, no massive brigandage, no piracy; we may travel at all hours and sail from East to West." An inscription from Hierapolis in Phrygia dating from the imperial period tells how an operator named Flavius Zeuxis passed Cape Taïnaron no fewer than 72 times.

The economy of mainland Greece declined during the Hellenistic age, though standards rose briefly about 260 BC, and there were pockets of prosperity, such as the Boeotian city of Tanagra famous for its terra-cotta figurines. The general picture is one of poverty, unemployment, falling wages, depopulation, and emigration. The forests were stripped, the land neglected, and smallholdings swallowed up in large estates, which, however, were underdeveloped. The Athenian silver mines at Laurium were depleted, though they reopened briefly at the end of the 3rd century BC. Demand for fine painted pottery had ceased. Athenian wine was of poor quality. Olive oil, however, continued to command a market, so much so that a law of AD 125 reserved one-third of the production to indigenous use; but, as the historian Moses I. Finley argued, olive oil alone would hardly meet the balance of payments. The centres of Hellenic prosperity had shifted with the movement of Hellenism from Athens, Corinth, Sparta, and Argos to Alexandria, Rhodes, Pergamum, and Antioch.

Trade in the Mediterranean basin

Within the Mediterranean basin, trade was mostly in essentials or things regarded as such. Metals ranked highest in importance: there was silver from Spain, copper from Cyprus, iron from the Black Sea coast and later China, and tin from Cornwall. Food also was important: grain came from Egypt, North Africa, the Crimea, and perhaps Babylonia. In other areas there was some specialization: Athens was noted for honey as well as olive oil, Byzantium for fish, Jericho for dates, and Damascus for prunes. Textiles were prominent: linen arrived from Egypt, a kind of silk from Tyre and true silk from China, and woolen goods from Miletus. Timber came from the forests of Anatolia and the north, marble for building from Paros and Athens, granite from Egypt: some docks constructed in Delos about 130 BC are of Egyptian granite.

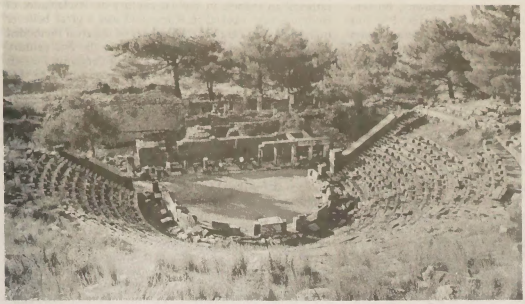
The prosperity of Egypt, "the gift of the Nile," was rooted in agriculture. The land lent itself to the cultivation of wheat, barley and sorghum, flax, vegetables (including lentils, beans, chickpeas, and onions), the date palm, and papyrus, as well as the raising of animals, such as horses, donkeys, goats, cattle, poultry, and fish.

Strabo gives a vivid picture of the resources of the Seleucid kingdom. He speaks of the rich yields of barley and the varied uses of the products of the palm—for food, drink, sweetening, fuel, and weaving. Mesopotamia is "good pastureland, and rich in vegetation, evergreens and spice." Rice was introduced into Persia from India, and the vine from Greece.

Similarly Strabo identifies the specialties of different regions of Anatolia. He mentions the fruit trees, vines, and olives of Melitene; the stone, timber, and pastures around Mazaca; the orchards of Cappadocia, and its mineral resources in red ochre, crystal, onyx, and mica; the market gardens of Sinope and beyond them olive groves and timber forests; the cattle and cheese of Bithynia; the *styrax*, producing gum and wood for spears, of the Taurus Mountains; and the superb wools of Laodicea and Colossae.

One figure suffices to indicate the huge economic expansion during the Hellenistic age. The customs revenue of Rhodes in about 170 BC was five times that of Athens in 400, with almost certainly the identical rate of 2 percent. It would be hard to demonstrate more clearly that the Hellenistic world operated in a totally different dimension.

Cultural developments. *Architecture.* It was in the Hellenistic age that the grid plan came into its own, in the



The Hellenistic theatre at Priene, second half of the 4th century BC.
© Jason Wood Photographs

numerous new foundations, and some of the refoundations, such as Priene.

The great buildings of the Classical age had been predominantly religious; those of the Hellenistic age were predominantly secular, though it will not do in the ancient world to make a rigid distinction between the two. The chief characteristic of Hellenistic temple architecture was the predilection for the Corinthian style, which came into its own with the Temple of Olympian Zeus at Athens, begun in 174 BC. Many Hellenistic temples were of immense size; this one is 135 feet by 354 feet on the stylobate. The oracular temple of Apollo at Didyma is 168 feet by 359 feet on the stylobate. Another colossal temple was built at Cyzicus in the 2nd century AD, with columns of more than 6.5 feet in diameter; it displaced the temple of Artemis at Ephesus as one of the Seven Wonders of the World.

Theatres

Some of the theatres were similarly colossal. Hieron II's 3rd-century modifications of the rock-cut theatre in Syracuse and the theatres at Megalopolis and Ephesus accommodated more than 20,000 people. There were changes of design, initiated at Athens with the emergence of New Comedy, which eliminated the chorus from a significant part in the drama. The result was the introduction of a high shallow stage, removable for revivals of the ancient plays and therefore of wood. Later the proscenium was built in from the first, and eventually it was constructed of stone, as at Oropus in about 200 BC; at Athens the change was deferred until about 150 BC. The Roman-built theatres are distinguishable by the fact that the auditorium is a perfect semicircle. The orchestra was often expanded for gladiatorial and wild animal fights and correspondingly surrounded by a high wall; at Stobi, Tyndaris, and Corinth this was more than 9.8 feet in height. Roman theatres were often built standing free rather than fitted into a hillside: the magnificent theatre at Aspendus is an example. The best-preserved theatre of the Roman Hellenistic world is at Bostra Traiani in present-day Syria.

All stadiums by definition ought to have the course of a given length, though, curiously, they vary by more than 30 feet. The stadium at Athens was built in the shape of a U with one flat end and one rounded; it was reconstructed in Pentelic marble in AD 143 by the millionaire benefactor Herodes Atticus. The great stadiums at Aphrodisias and Nysa in Anatolia and at Laodicea in Syria belong to the Roman period and are rounded at both ends. The one at Aphrodisias seated 30,000 people and is excellently preserved. The gymnasium and palaestra tended in the Hellenistic period to be more formalized in plan and structure.

Houses

The palaces of the Greek period have not survived. Remaining houses show increasing elaboration and luxury. Examples from the 3rd century may be seen in Priene, consisting normally of a block of four rooms with a pillared entrance opening on to a courtyard. In some of the wealthier houses, rooms are found on three sides of the court, and there may be columns opening onto an entrance corridor on the fourth. This structure developed into a peristyle house already found in Olynthus in the 4th century. Delos has a variety of peristyle houses built on irregular plans; generally one finds a great water cistern and often spectacular mosaics. In southern Italy the Greek population developed its own style of house, whose court in Pompeii blended with that of the peristyle structure. These houses presented to the street generally bare walls. The typical house is symmetrical about its long axis. A short hall reaches an atrium, or lofty court with an impluvium, or cistern, at its centre.

The arts. Hellenistic sculpture, often of a very high quality, is notable for its variety. Alexander's *pothos,* or yearning for something unattained, was a mood that became expressed in the art. Lysippus, Alexander's favourite sculptor, had produced a seminal statue, the "Apoxyomenos" ("The Athlete, Scraping Himself"), a figure standing with one arm extended and the other pulled across his body. The viewer has to move around it because no single viewpoint is satisfactory. Eutychides, a pupil of Lysippus, carried the principle further in his portrayal of "The Fortune of Antioch." Vastly more complex, and showing the search for an original subject, is the brilliant and brutal "The Punishment of Dirce" by Apollonius and Tauriscus of Tralles. "Laocoön," a portrayal of anguish, shows the figure of the priest Laocoön and his two sons in the grip of two snakes. The sculpture, in immobile stone, is bursting with dynamism and energy.

Pergamum was one of the great centres of sculpture. There Attalus I commemorated his victory over the Gauls with a huge monumental group on a circular base. The altar of Zeus at Pergamum bore a frieze 364 feet long portraying the battle of the gods and giants; muscular superhuman figures are rendered in dynamic, agonized conflict.

Genre scenes in art

An aspect of the Hellenistic search for variety was the use of the genre subject, such as a boy with a goose, a drunken old hag, a boy pulling a thorn from his foot. The attractive terra-cotta figurines from Tanagra and Myrina offer a fine selection of scenes from ordinary life, such as a grossly fat nurse with a bulbous nose holding a baby in her lap, a boy wearing a dunce's cap, two women gossiping, or acrobats in all manner of attitudes. The search for variety, paradoxically, also took the form of a return to the Classical style. Examples are the "Venus de Milo," whose face recalls the manner of the 4th-century sculptor Praxiteles, and the "Belvedere Torso," modeled on a 4th-century sculpture but with a muscular twist that marks it as Hellenistic.

Portraiture was a natural accompaniment of the courts. Rulers were finely portrayed not just in statues but on coins. Some of the finest of these come from the outlying kingdoms of Bactria and India. The portraits do not always

flatter; the monarchs appear podgy or scrawny, broken-nosed or hook-nosed. Full statues were rarer. Portraits were not confined to rulers. The statue of Demosthenes in Copenhagen, taut and intense, is copied from a 3rd-century original by Polyeuctus, sculpted well after the orator's death. Philosophers were often depicted; although it is possible to distinguish individuals, a type of "philosopher" is imposed on them.

Literature. In literature, just as in the arts, one finds a combination of novelty and commonplace types and themes. In the New Comedy at Athens, of which Menander (*c.* 342–292 BC) was the leading exponent, the theme is no longer fantasy but real life. The plays are not uproarious, as those of Aristophanes can be, but they are filled with quiet good humour. Besides Menander, there was Herodas (3rd century BC), who in his *Mimiambi* (*Mimes*) sketched episodes from life. Theophrastus (*c.* 370–287 BC) produced a minor masterpiece, *Characters,* in which he depicted such figures as the Stupid Man, who cannot remember where he lives, or the Tactless Man, who makes a misogynistic speech at a wedding.

The New Comedy

Some writers took a deeper interest in psychology. The poet Apollonius of Rhodes (3rd century BC) wrote an epic on the Argonauts, in which he closely observed the psychology of Medea at her first experience of love; his sensitive and romantic rendition influenced the Roman poet Virgil in his portrayal of the ill-fated love between Dido and Aeneas. Theocritus (*c.* 300–260 BC), who came from Sicily but lived mostly in Cos and Alexandria, examined in his second idyll the love-hate relationship of a girl to her unfaithful lover. The world of Theocritus is a world of pastoral artifice having little to do with the real hardships of country life, but the details are exquisitely noticed.

Alexandria was noted for its learning. The poet Callimachus (*c.* 305–240 BC), who was attached to the city's famous library, wrote poetry of polished craft and allusive scholarship. His great work *Aetia* ("causes") is a rare miscellany, a long poem made up of short sections. Callimachus, immensely influential, has quality.

The major contributions to prose literature fall in the Roman period, though the novel developed earlier in Alexandria. Ingenious and exciting plots are combined with stereotyped characters. Longus' *Daphnis and Chloe* (date unknown) is perhaps the best of such works of prose fiction. Another important development was the rhetoric of the movement known as the Second Sophistic, which belongs mainly to the 2nd century AD. Its finest practitioner was Dion of Prusa (*c.* AD 40–112), nicknamed Chrysostom. Herodes Atticus (*c.* AD 101–177) and the flowery Polemo (*c.* AD 88–144) had much influence; more survives from the dull, Athens-loving hypochondriac Aelius Aristides (*c.* AD 117–187) and the facile Maximus of Tyre (*c.* AD 125–185). Greater than any of these is the Syrian Lucian (*c.* AD 120–185), a satirist and brilliant entertainer, who spared neither gods nor humans.

Other writers, worthy enough, must receive passing mention: they are the geographers Strabo (*c.* 64 BC–AD 25) and Ptolemy and Pausanias (both 2nd century AD), the historians Diodorus Siculus of Sicily (1st century BC), Arrian (2nd century AD), Appian (2nd century AD) and Dio Cassius (2nd–3rd century AD), the voluminous Jewish writers Philo Judaeus and Flavius Josephus (*c.* AD 37–100), the vastly miscellaneous Athenaeus (*c.* AD 200), the historian and teacher of rhetoric Dionysius of Halicarnassus (fl. late 1st century BC), and the unknown writer (conventionally known as Longinus) of a major work *On the Sublime* (1st century AD), with his acute observations about Homer and Sappho, Demosthenes and Thucydides, and even about "the Jewish lawgiver" in *Genesis.*

Science and medicine. The three great areas of Hellenistic scholarship were medicine, astronomy, and mathematics. Alexandria attracted Herophilus (fl. 3rd century BC) from Chalcedon, who refused to stand in awe of the accepted medical dogmas and was distinguished in systematic anatomy, and the notable physiologist Erasistratus (fl. 3rd century BC) from Ceos, who realized that the heart is the motor for the circulatory system and deduced the existence of capillaries. Philinus (fl. 3rd century BC) from Cos founded the empirical school, trusting clinical observation rather than theory. In the 1st century BC Asclepiades of Bithynia, who worked in Rome and was a great believer in hygiene, was claimed the founder of the rival methodist school, based on Epicurean atomism. In the 2nd century emerged the towering figure of Galen of Pergamum (*c.* AD 129–199), whose authority later was second only to that of Aristotle.

Medicine, astronomy, mathematics

In astronomy the first great advances were due to Aristarchus of Samos in the early 3rd century BC. He was the pioneer of the theory that the Sun is at the centre of the universe. His greatest achievement lay in his method for determining the sizes and distance of the Sun and the Moon, though his observational technique was inadequate for correct results. Later in the century Eratosthenes of Cyrene, a typical polymath, calculated the Earth's circumference by an excellent method, though his good result was due to the mutual canceling out of two errors.

In mathematics the key figures are Euclid (fl. *c.* 300 BC), Archimedes (*c.* 287–212 BC), and Apollonius (fl. late 3rd century BC). Euclid, whose *Elements* served as a basic textbook of geometry for 2,000 years, was both a systematizer and original mathematician. Archimedes preferred to concentrate on particular problems, working in the realms of geometry, physics, and mechanics, and he formulated the science of hydrostatics. Apollonius of Perga was the great authority on conics. One other significant mathematician was Hero of Alexandria (fl. 1st century AD), who actually devised a simple steam engine but treated it as a mere toy.

Philosophy. The philosophers of the period pursued *autarkeia:* self-sufficiency, or nonattachment. The most extreme position was taken by the cynics, whose founder was Diogenes of Sinope (*c.* 400–325 BC). Behind his rejection of traditional allegiances lay a profound concern with moral values. What matters to human beings, he taught, was not social status or nationality but individual well-being, achieved by a reliance on one's natural endowments. He was followed by the attractive couple Crates (*c.* 365–285 BC) and Hipparchia. Zeno (335–263 BC), founder of the stoics, began from here. To the stoics nothing is good but virtue, nothing bad but vice; all else is indifferent. The stoics were pantheists. They believed that all is in the hands of God; indeed, God is all. Moreover, all is for the best in the best of all possible worlds, and human beings only have to accept and give praise. Zeno was succeeded by a religious genius named Cleanthes (331–232 BC) and he by the great systematizer Chrysippus (*c.* 280–207 BC). The 2nd century produced Panaetius (*c.* 185–109 BC), who smoothed away some of the sharper stoic paradoxes for the Romans, and the 1st brought Poseidonius (*c.* 135–50 BC), another mediator between east and west.

Cynics and stoics

Epicurus (341–270 BC), an Athenian contemporary of Zeno, stood poles apart in thought from the stoics. In opposition to their moralism he taught that the goal of life is pleasure, a position for which he has been much maligned. In fact, he advocated the simple life as being the most pleasurable and said that it was impossible to live pleasurably without being wise, just, and honest. (J.Fe.)

Ancient Italic peoples

Pre-Roman Italy was inhabited by peoples diverse in origin, language, traditions, stage of development, and territorial extension and was heavily influenced by neighbouring Greece, with its well-defined national characteristics, expansive vigour, and aesthetic and intellectual maturity. Italy attained a unified ethnolinguistic, political, and cultural physiognomy only after the Roman conquest; yet its most ancient peoples remain anchored in the names of the regions of Roman Italy—Latium, Campania, Apulia, Bruttium, Lucania, Samnium, Picenum, Umbria, Etruria, Venetia, and Liguria.

THE ETRUSCANS

The Etruscans formed the most powerful nation in pre-Roman Italy. They created the first great civilization on the peninsula, whose influence on the Romans as well as on 20th-century culture is increasingly recognized. Evidence suggests that it was the Etruscans who taught the Romans the alphabet and numerals, along with many el-

ements of architecture, art, religion, and dress. The toga was an Etruscan invention, and the Etruscan-style Doric column (rather than the Greek version) became a mainstay of architecture of both the Renaissance and the later Classical revival. Etruscan influence on the ancient theatre survives in their word for "masked man," *phersu,* which became *persona* in Latin and *person* in English.

General considerations. *Nomenclature.* The Greeks called the Etruscans Tyrsenoi or Tyrrhenoi, while the Latins referred to them as Tusci or Etrusci, whence the English name for them. In Latin their country was Tuscia or Etruria. According to the Greek historian Dionysius of Halicarnassus (fl. *c.* 20 BC), the Etruscans called themselves Rasenna, and this statement finds confirmation in the form *rasna* in Etruscan inscriptions.

"Rasenna"

Geography and natural resources. Ancient Etruria lay in central Italy, bounded on the west by the Tyrrhenian Sea (recognized early by the Greeks as belonging to the Tyrrhenoi), on the north by the Arno River, and on the east and south by the Tiber River. This area corresponds to a large part of modern Tuscany as well as to sections of Latium and Umbria. The chief natural resources of the region, undoubtedly playing a crucial role in Etruscan commerce and urban development, were the rich deposits of metal ores found in both northern and southern Etruria. In the south, in the maritime territory stretching between the first great Etruscan cities, Tarquinii and Caere (modern Cerveteri), the low-lying Tolfa Mountains provided copper, iron, and tin. These minerals also were found inland at Mount Amiata, the highest mountain in Etruria, in the vicinity of the city of Clusium (modern Chiusi). But the most productive area turned out to be in northern Etruria, in the range known as the Catena Metallifera ("Metal-Bearing Chain"), from which copper and especially iron were mined in enormous amounts. The city of Populonia, located on the coast, played a leading role in this industry, as did the adjacent island of Elba, evidently renowned from an early date for the wealth of its deposits.

The forests of Etruria constituted another major natural resource, providing abundant firewood for metallurgical operations as well as timber for the building of ships. The Etruscans were famous, or perhaps infamous, for their maritime activity; they dominated the seas on the western coast of Italy, and their reputation as pirates instilled fear around the Mediterranean. Their prosperity through the centuries, however, seems also to have been founded on a stout agricultural tradition; as late as 205 BC, when Scipio Africanus was outfitting an expedition against Hannibal, the Etruscan cities were able to supply impressive amounts of grain as well as weapons and materials for shipbuilding.

Historical periods. The presence of the Etruscan people in Etruria is attested by their own inscriptions, dated about 700 BC; it is widely believed, however, that the Etruscans were present in Italy before this time and that the prehistoric Iron Age culture called "Villanovan" (9th–8th century BC) is actually an early phase of Etruscan civilization.

Inasmuch as no Etruscan literary works have survived, the chronology of Etruscan history and civilization has been constructed on the basis of evidence, both archaeological and literary, from the better-known civilizations of Greece and Rome as well as from those of Egypt and the Middle East. Contact with Greece began around the time that the first Greek colony in Italy was founded (*c.* 775–750 BC), when Greeks from the island of Euboea settled at Pithekoussai in the Bay of Naples. Thereafter, numerous Greek and Middle Eastern objects were imported into Etruria, and these items, together with Etruscan artifacts and works of art displaying Greek or Oriental influence, have been used to generate relatively precise dates along with more general ones. In fact, the basic nomenclature for the historical periods in Etruria is borrowed from corresponding periods in Greece; the assigned dates are usually (though perhaps erroneously) conceived of as being slightly later than their Greek counterparts to allow for cultural "time lag." Thus the Etruscan Orientalizing period belongs to the 7th century BC; the Archaic period to the 6th and first half of the 5th century BC; the Classical period to the second half of the 5th and the 4th century BC; and the Hellenistic period to the 3rd to 1st centuries BC. Etruscan culture became absorbed into Roman civilization during the 1st century BC and thereafter disappeared as a recognizable entity.

Language and writing. Etruscan, the third great language of culture in Italy after Greek and Latin, does not, as noted above, survive in any literary works. An Etruscan religious literature did exist, and evidence suggests that there may have been a body of historical literature and drama as well. (Known, for example, is the name of a playwright, Volnius, of obscure date, who wrote "Tuscan tragedies.") Etruscan had ceased to be spoken in the time of imperial Rome, though it continued to be studied by priests and scholars. The emperor Claudius (d. AD 54) wrote a history of the Etruscans in 20 books, now lost, which was based on sources still preserved in his day. The language continued to be used in a religious context until late antiquity; the final record of such use relates to the invasion of Rome by Alaric, chief of the Visigoths, in AD 410, when Etruscan priests were summoned to conjure lightning against the barbarians.

Etruscan inscriptions

There are more than 10,000 known Etruscan inscriptions, with new ones being discovered each year. These are mainly short funerary or dedicatory inscriptions, found on ash urns and in tombs or on objects dedicated in sanctuaries. Others are found on engraved bronze Etruscan mirrors, where they label mythological figures or give the name of the owner, and on coins, dice, and pottery. Finally, there are graffiti scratched on pottery; though their function is little understood, they seem to include owners' names as well as numbers, abbreviations, and nonalphabetic signs.

The Metropolitan Museum of Art, New York City, Fletcher Fund, 1924 (24.97.21ab)



Bucchero ware jug in the shape of a cock, incised with the Etruscan alphabet on both sides, *c.* 7th–8th century BC, from Viterbo. In the Metropolitan Museum of Art, New York City.

Of the longer inscriptions, the most important is the "Zagreb mummy wrapping," found in Egypt in the 19th century and carried back to Yugoslavia by a traveler (National Museum, Zagreb). It had originally been a book of linen cloth, which at some date was cut up into strips to be wrapped around a mummy. With about 1,300 words, written in black ink on the linen, it is the longest existing Etruscan text; it contains a calendar and instructions for sacrifice, sufficient to give some idea of Etruscan religious literature. From Italy come an important religious text, inscribed on a tile at the site of ancient Capua, and an inscription on a boundary stone at Perugia, noteworthy for its juridical content. The few Etruscan-Latin bilingual inscriptions, all funerary, have little importance with respect to improving knowledge of Etruscan. But inscribed gold plaques found at the site of the ancient sanctuary of Pyrgi, the port city of Caere, provide two texts, one in Etruscan and the other in Phoenician, of significant length (about 40 words) and of analogous content. They are the equivalent of a bilingual inscription and thus offer substantial data for the elucidation of Etruscan by way of

a known language—Phoenician. The find is also an important historical document, which records the dedication to the Phoenician goddess Astarte of a "sacred place" in the Etruscan sanctuary of Pyrgi by Thefarie Velianas, king of Caere, early in the 5th century BC.

The 20th-century notion that there is a "mystery" regarding the Etruscan language is fundamentally erroneous; there exists no problem of decipherment, as is often wrongly asserted. The Etruscan texts are largely legible.

Etruscan alphabet

The alphabet derives from a Greek alphabet originally learned from the Phoenicians. It was disseminated in Italy by the colonists from the island of Euboea during the 8th century BC and adapted to Etruscan phonetics; the Latin alphabet was ultimately derived from it. (In its turn the Etruscan alphabet was diffused at the end of the Archaic period [*c.* 500 BC] into northern Italy, becoming the model for the alphabets of the Veneti and of various Alpine populations; this happened concurrently with the formation of the Umbrian and the Oscan alphabets in the peninsula.)

The real problem with the Etruscan texts lies in the difficulty of understanding the meaning of the words and grammatical forms. A fundamental obstacle stems from the fact that no other known language has close enough kinship to Etruscan to allow a reliable, comprehensive, and conclusive comparison. The apparent isolation of the Etruscan language had already been noted by the ancients; it is confirmed by repeated and vain attempts of modern science to assign it to one of the various linguistic groups or types of the Mediterranean and Eurasian world. However, there are in fact connections with Indo-European languages, particularly with the Italic languages, and also with more or less known non-Indo-European languages of western Asia and the Caucasus, the Aegean, Italy, and the Alpine zone as well as with the relics of the Mediterranean linguistic substrata revealed by place-names. This means that Etruscan is not truly isolated; its roots are intertwined with those of other recognizable linguistic formations within a geographic area extending from western Asia to east-central Europe and the central Mediterranean, and its latest formative developments may have taken place in more direct contact with the pre-Indo-European and Indo-European linguistic environment of Italy. But this also means that Etruscan, as scholars know it, cannot simply be classified as belonging to the Caucasian, the Anatolian, or Indo-European languages such as Greek and Latin, from which it seems to differ in structure.

Methods of interpreting the Etruscan language

The traditional methods hitherto employed in interpreting Etruscan are (1) the etymological, which is based upon the comparison of word roots and grammatical elements with those of other languages and which assumes the existence of a linguistic relationship that permits an explication of Etruscan from the outside (this method has produced negative results, given the error in the assumption), (2) the combinatory, a procedure of analysis and interpretation of the Etruscan texts rigorously limited to internal comparative study of the texts themselves and of the grammatical forms of the Etruscan words (this has led to much progress in the knowledge of Etruscan, but its defects lie in the hypothetical character of many of the conclusions due to the absence of external proofs or confirmations), and (3) the bilingual, based on the comparison of Etruscan ritual, votive, and funerary formulas with presumably analogous formulas from epigraphic or literary texts in languages belonging to a closely connected geographic and historical environment, such as Greek, Latin, or Umbrian. Nonetheless, with the increase of reliable data, in part from more recent epigraphic discoveries (such as the gold plaques at Pyrgi mentioned above), the need to find the one right method appears to be of decreasing importance; all available procedures tend to be utilized.

Archaeological evidence. The lack of Etruscan literature and the widely acknowledged bias and contradictory accounts of Greek and Roman writers create a situation in which the careful study of the visible remains of the Etruscans is fundamental for understanding them. The archaeological contexts and the remains themselves (including pottery, metalwork, sculpture, painting, architecture, animal and human bones, and the humblest objects of daily life) fall into three basic categories: funerary, urban, and sacred. (There is sometimes an overlapping of these categories.)

By far the largest percentage of material is funerary; thus there is a great deal of information about Etruscan ideas on the afterlife and on their attitudes toward the deceased members of their families. But there can be no doubt that the relatively scarce information about Etruscan settlements is also of great importance. The evidence of the well-preserved Etruscan city at Marzabotto (*c.* 500 BC) near Bologna (probably an Etruscan colony) reveals that the Etruscans were among the first in the Mediterranean to lay out a city with a grid plan; it was oriented according to the compass, emphasizing a principal north-south street and including one or more major east-west streets. The ritual involved in thus laying out a town, complete with walls, temples, and other sacred areas, was known to the Romans as the *ritus etruscus.* The system was commonly used by the Romans in laying out military camps and new cities and has survived in the centre of many European cities today. Such rigidly organized town plans seem to have been rare in Etruria; more often one finds an irregular pattern resulting from the coalescence of villages in Villanovan times and the adaptation to the hills normally chosen as town sites.

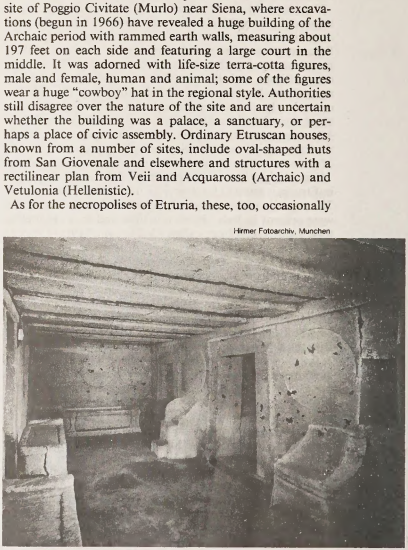
Settlement patterns

In a sacred context, the Etruscan temple also often revealed a careful organization, once again with a system that was passed on to the Romans. In contrast to Greek temples, those of the Etruscans frequently showed a clear differentiation of front and back, with a columniated deep front porch and a cella that was flush with the podium on which it stood. The materials were frequently perishable (timber and mud brick, on a stone foundation) except for the abundant terra-cotta sculptures that adorned the roof. Especially well-preserved are the acroteria, or roof sculptures, from the Portonaccio temple at Veii (late 6th century BC) representing Apulu (the Etruscan Apollo) and other mythological figures.

Of a different order are the spectacular finds from the site of Poggio Civitate (Murlo) near Siena, where excavations (begun in 1966) have revealed a huge building of the Archaic period with rammed earth walls, measuring about 197 feet on each side and featuring a large court in the middle. It was adorned with life-size terra-cotta figures, male and female, human and animal; some of the figures wear a huge "cowboy" hat in the regional style. Authorities still disagree over the nature of the site and are uncertain whether the building was a palace, a sanctuary, or perhaps a place of civic assembly. Ordinary Etruscan houses, known from a number of sites, include oval-shaped huts from San Giovenale and elsewhere and structures with a rectilinear plan from Veii and Acquarossa (Archaic) and Vetulonia (Hellenistic).

As for the necropolises of Etruria, these, too, occasionally

Hirmer Fotoarchiv, Munchen



Tomb of the Shields and Chairs, second half of 6th century BC, Caere, Italy.

show signs of a grid plan, as at the Crocefisso del Tufo at Orvieto (second half of the 6th century BC) and at Caere. More often they have an irregular, agglutinative quality that reflects the site's long history of use. Because the Etruscans took great pains to make their relatives comfortable in a "house of the dead," the tombs suggest many details of actual Etruscan houses. Thus the tombs of Caere (especially those of the 6th century and later), carved underground out of the soft volcanic tufa so widespread in Etruria, have not only windows, doors, columns, and ceiling beams but also pieces of furniture (beds, chairs, and footstools) sculptured from the living rock. At Tarquinii, another tradition for tomb decoration led to painting the walls of the chamber with frescoes of Etruscan funerary celebrations, including banqueting, games, dancing, music, and various performances in a fresh outdoor landscape. The scenes probably served to commemorate actual funerals, but they also may have alluded to the kind of afterlife that was expected for the deceased. The Elysium-like concept of the afterlife prevailed in the Archaic period, but in the ensuing centuries one finds a growing emphasis on the darker realm of the underworld. Frescoes show its ruler, Hades (Etruscan Aita), wearing a wolf-skin cap and sitting enthroned beside his wife; demons and monsters populate this sphere. They may be seen in the remarkable Tomb of the Blue Demons (*c.* 400 BC), discovered at Tarquinii in 1987, or in the Francois Tomb from Vulci, where the blue-skinned devil Charu (only remotely resembling the Greek ferryman Charon) waits with his hammer to strike the deceased and take him away to the underworld. He sometimes has a gentler partner, the angelic winged figure of Vanth, who helps to ease the transition from life to death.

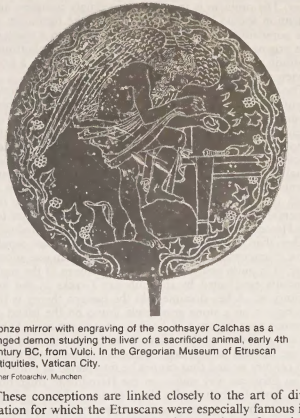
A perennial theme in the discussion of Etruscan material culture is its relationship to Greek models. The comparison is natural, indeed essential, in light of the massive amount of Greek artifacts, especially vases, that have been excavated in Etruria and the abundant examples of Etruscan imitations, of the pottery especially. It is also certain that Greek craftsmen sometimes settled in Etruria, as in the report by Pliny the Elder (1st century AD) about a Corinthian noble named Demaratus, who moved to Tarquinii, bringing along three of his own artists. But it is no longer appropriate to dwell naively on the "inferiority" of Etruscan art nor to insist that the Etruscans were mere imitators of the Greek art they undoubtedly prized. Instead, increasing emphasis is being placed on defining the highly original elements in Etruscan culture that exist side by side with the qualities that show their great admiration of things Greek.

Relationship between Greek and Etruscan culture

In addition to their distinctive modes of designing a town or of building a temple or tomb, one may note their unique native pottery, bucchero (beginning *c.* 680 BC), with its decorative incision in a shiny black fabric; it is radically different from standard Greek vase decoration, which regularly featured paint and a contrast of red or cream and black. In metallurgy, their bronze mirrors, sometimes described as an Etruscan "national industry," featured a convex reflecting side and a concave side adorned with engravings of themes from Greek and Etruscan mythology and daily life. Etruscan fashion also had many unique elements such as a hem-length braid down the back (7th century BC), pointed-toe shoes (*c.* 575–475 BC), and the mantle with the curved hem known to the Romans as the toga (6th century BC and later). Finally, the Etruscans seem to have taken an early interest in reproducing the features of their honoured relatives or officials (as in the funerary canopic urns from Clusium) and thus gave a major impetus to the development of truly realistic portraiture in Italy (especially in the Hellenistic period).

Religion and mythology. The essential ingredient in Etruscan religion was a belief that human life was but one small meaningful element in a universe controlled by gods who manifested their nature and their will in every facet of the natural world as well as in objects created by humans. This belief permeates the Etruscan representational arts, where one finds rich depictions of land, sea, and air, with man integrated into the ambient. Roman writers give repeated evidence that the Etruscans regarded every bird and every berry as a potential source of knowledge of the gods and that they had developed an elaborate lore and attendant rituals for using this knowledge. Their own myths explained the lore as having been communicated by the gods through a prophet, Tages, a miraculous child with the features of a wise old man who sprang from a plowed furrow in the fields of Tarquinii and sang out the elements of what the Romans called the *Etrusca disciplina.*

The literary, epigraphic, and monumental sources provide a glimpse of a cosmology whose image of the sky with its subdivisions is reflected in consecrated areas and even in the viscera of animals. The concept of a sacred space or area reserved for a particular deity or purpose was fundamental, as was the corollary theory that such designated areas could correspond to each other. Heaven reflected Earth, and macrocosm echoed microcosm. The celestial dome was divided into 16 compartments inhabited by the various divinities: major gods to the east, astral and terrestrial divine beings to the south, infernal and inauspicious beings to the west, and the most powerful and mysterious gods of destiny to the north. The deities manifested themselves by means of natural phenomena, principally by lightning. They also revealed themselves in the microcosm of the liver of animals (typical is a bronze model of a sheep's liver found near Piacenza, bearing the incised names of divinities in its 16 outside divisions and in its internal divisions).



Bronze mirror with engraving of the soothsayer Calchas as a winged demon studying the liver of a sacrificed animal, early 4th century BC, from Vulci. In the Gregorian Museum of Etruscan Antiquities, Vatican City.
Hirmer Fotoarchiv, Munchen

These conceptions are linked closely to the art of divination for which the Etruscans were especially famous in the ancient world. Public and private actions of any importance were undertaken only after having interrogated the gods; negative or threatening responses necessitated complex preventive or protective ceremonies. The most important form of divination was haruspicy, or hepatoscopy—the study of the details of the viscera, especially the livers, of sacrificial animals. Second in importance was the observation of lightning and of such other celestial phenomena as the flight of birds (also important in the religion of the Umbri and of the Romans). Finally, there was the interpretation of prodigies—extraordinary and marvelous events observed in the sky or on the earth. These practices, extensively adopted by the Romans, are explicitly attributed by the ancient authors to the religion of the Etruscans.

Divination

The Etruscans recognized numerous deities (the Piacenza liver lists more than 40), and many are unknown today. Their nature was often vague, and references to them are fraught with ambiguity about number, attributes, and even gender. Some of the leading gods were eventually equated with major deities of the Greeks and Romans,

Etruscan deities

as may be seen especially from the labeled representations on Etruscan mirrors. Tin or Tinia was equivalent to Zeus/Jupiter, Uni to Hera/Juno, Sethlans to Hephaestus/Vulcan, Turms to Hermes/Mercury, Turan to Aphrodite/Venus, and Menrva to Athena/ Minerva. But their character and mythology often differed sharply from that of their Greek counterparts. Menrva, for example, an immensely popular deity, was regarded as a sponsor of marriage and childbirth, in contrast to the virgin Athena, who was much more concerned with the affairs of males. Many of the gods had healing powers, and many of them had the authority to hurl a thunderbolt. There were also deities of a fairly orthodox Greco-Roman character, such as Hercle (Heracles) and Apulu (Apollo), who were evidently introduced directly from Greece yet came to have their designated spaces and cults.

Origins. Because the Etruscans spoke a non-Indo-European language while being surrounded in historical times by Indo-European peoples such as the Latins and Umbro-Sabelli, scholars of the 19th century examined and debated, often bitterly, the origins of this anomalous population. Their dispute continued into the 20th century but has now lost much of its intensity. The leading scholar in Etruscan studies, Massimo Pallottino, has wisely observed that such discussions have become sterile as the result of an incorrect formulation of the problem. Too much emphasis has been placed on the provenance of the Etruscans, with the expectation that there could be one simple answer. The problem is in reality exceedingly complex, and attention should be directed instead to the formation of the population, as it might be, for example, in a study of the origins of "the Italians" or "the French." Pallottino's position may be understood more clearly through a brief review of the debate.

The debate about the Etruscans' provenance

The argument began, in fact, in antiquity with the statement by Herodotus that the Etruscans migrated from Lydia in Anatolia shortly after the time of the Trojan War; their leader was Tyrsenos, who later gave his name to the whole race. Supporters of this "Eastern" theory pointed above all to the archaeological evidence of profound Oriental influence on Etruscan culture, such as in monumental funerary architecture and exotic luxury goods of gold, ivory, and other materials. But chronologically the Oriental inundation occurred nearly 500 years too late for the Herodotean migration. Further, it developed gradually rather than making the sudden appearance that would have characterized the arrival of a people en masse; moreover, it is quite easily explained by reference to the trade conduits established by the Euboean Greeks in the 8th century BC. A key document in the Eastern theory is the inscription on a stone grave stela found on the island of Lemnos near the coast of Anatolia that shows remarkable lexical and structural similarities with the Etruscan language. But this curious isolated document dates only to the 6th century BC and thus cannot be interpreted as evidence of an Etruscan way station in the Herodotean migration from Anatolia to Italy. On the contrary, it has now been proposed that Lemnos may in fact have been colonized or used as a trading point by the Etruscans looking toward Anatolia in the 6th century BC rather than as a place they visited moving away from the area.

A second theory on Etruscan origins was proposed by Dionysius of Halicarnassus, who rejected the tradition of Herodotus, pointing out that the Lydian language and customs and those of the Etruscans were greatly dissimilar; he argued that the Etruscans were autochthonous (of local origin). Acceptance of this "autochthonous" theory requires that Villanovan culture be regarded as an early phase of Etruscan civilization (a hypothesis now widely endorsed) and, in addition, that there be links with an ethnic substratum of the Bronze Age in Italy (2nd millennium BC). There are indeed stray affinities with the Bronze Age culture of the "Terramara," with its cremating, sedentary habits, but also with the "Apenninic" culture, which was seminomadic and practiced inhumation. There is, however, mounting evidence of a critical transition period at the end of the Bronze Age and the beginning of the Iron Age, in which there are so many important developments that the connections between these two cultures and the Villanovan seem minor. Although the terminology is vexed for this transition period, varying from "sub-Apennine" to "Recent Bronze," "Final Bronze," and, most frequently, "Proto-Villanovan," the social and economic changes are clear. There was an increase in population and in overall wealth, a tendency to have larger, permanent settlements, an expansion of metallurgical knowledge, and a strengthening of agricultural technology. Diagnostic archaeological criteria include the use of cremation (with a biconical ash urn) and the presence of characteristic artifacts such as the fibula ("safety pin"), razor, objects of amber, the ax, and various other bronze weapons. The fact that the Proto-Villanovan archaeological horizon developed gradually rather than suddenly as the result of invasion or large migration might seem to support the theory of autochthony for the Etruscans. But once again the picture is clouded, because the Proto-Villanovan occurs in scattered areas all around Italy, including zones that definitely did not emerge as Etruscan in historical times.

A 19th-century theory

To these two theories from antiquity was added a third in the 19th century to the effect that the Etruscans migrated overland into Italy from the north. This theory, without any ancient literary support, was based on similarities in customs and artifacts between the Villanovan and the Iron Age cremating cultures north of the Alps and on a dubious comparison of the name of the Rasenna with that of the Raeti, a people inhabiting the east-central Alps in the 5th century BC. The theory is basically without supporters today, though the influence or presence of certain central European weapon and helmet types and vessel forms in Etruria is not denied. These elements, however, are now put into perspective as representing simply one significant strand in the complex fabric of Etruscan culture as it developed from Villanovan to Orientalizing.

These northern connections in a sense form a parallel to the Greek influences in subsequent periods, whether Euboean (8th century BC), Corinthian (7th century), Ionian (6th century), or Attic (5th century). Likewise, Oriental influences may be readily acknowledged, coming from such diverse areas as Lydia, Urartu, Syria, Assyria, Phoenicia, and Egypt. But none of these connections per se give any firm proof about Etruscan "origins," and current scholarship is much more concerned with understanding the interrelationship of these influences and the context in which the civilization in Etruria developed.

Expansion and dominion. Archaeological evidence helps to develop a picture of the beginnings of Etruscan cities during the Villanovan period. Nearly every major Etruscan city of historical times has yielded Villanovan remains, but it is in the south, particularly near the coast, that the earliest signs of city formation appear. It is hypothesized that clusters of huts forming a network of villages on a single hill or on several adjacent hills coalesced into preurban settlements at this time. (The plural form of the names of some of these—Vulci, Tarquinii, and Veii—is consistent with this hypothesis.) Ash urns in the shape of oval huts with thatched roofs excavated in the area suggest what the houses of the living may have looked like, while the parity of grave goods for men and women implies a basically egalitarian society, at least in earlier stages. Cremation with ashes in a biconical vessel is commonly found as a holdover from the Proto-Villanovan; inhumation also appeared and during the Orientalizing period eventually became the prevailing rite, except in northern Etruria, where cremation persisted to the 1st century BC.

The Orientalizing period

After contact was made with Greeks and Phoenicians, new ideas, materials, and technology began to appear in Etruria. In the Orientalizing period the use of writing, the potter's wheel, and monumental funerary architecture accompanied the accumulation of luxury goods of gold and ivory and exotic trade items such as ostrich eggs, tridacna shells, and faience. The Regolini-Galassi Tomb at Caere (*c.* 650–625 BC), discovered in 1836 in an unplundered state, dramatically revealed the full splendour of the Orientalizing period. The tomb's main chamber belonged to a fabulously wealthy lady who, inhumed with her banquet service and a wide array of jewelry made by granulation and repoussé, might well be called a queen; the word *Larthia* on her belongings may record her name. Even if

Caere did not have kings and queens at this time (as did Rome, or as Caere certainly did in the 5th century), it is clear that society had become sharply differentiated, not only in regard to wealth but also in division of labour. Many scholars hypothesize the existence of a powerful aristocratic class, and craftsmen, merchants, and seamen would have formed a middle class; it was probably at this time that the Etruscans began to maintain the elegant slaves for which they were famous. (Various Greek and Roman authors report on how Etruscan slaves dressed well and how they often owned their own homes. They easily became liberated and rapidly rose in status once they were freed.)

The dramatic growth of Etruscan civilization and influence in the 7th century is reflected in the so-called "princely" tombs, closely akin to the Regolini-Galassi Tomb, found in Etruria itself at Tarquinii, Vetulonia, and Populonia and along the Arno River (*e.g.*, at Quinto Fiorentino) and in the south at Praeneste in Latium and at Capua and Pontecagnano in Campania. Literary sources report that Rome itself came under the rule of Etruscan kings in the late 7th century. Livy describes the arrival from Tarquinii of Tarquinius Priscus, the later king, and his ambitious, learned wife Tanaquil, a worthy counterpart to Queen Larthia of Caere. There is also archaeological evidence of Etruscan expansion northward into the Po valley in the 6th century.

True urbanization followed these developments. Mighty city-states featuring fortified walls and other public works flourished both in Etruria and in its spheres of influence. The Rome of the Etruscan kings, described in detail by Livy and known through excavation, had fortifications, a paved forum, a master drainage system (the Cloaca Maxima), a public stadium (the Circus Maximus), and a monumental Etruscan-style temple dedicated to Jupiter Optimus Maximus.

It is at the end of the 6th century that one finds the earliest evidence for the grid system in towns and cemeteries mentioned earlier. The ample but surprisingly uniform houses and tombs imply growing regulation and cooperation and possibly signal a change in government. Etruscan cities, like Rome itself, may have begun to remove their kings at this time and to operate under an oligarchic system with elected officials from powerful noble families.

The Roman orator Cato's statement that "almost all of Italy was once under Etruscan control" best applies to this period. Undoubtedly, Etruscan maritime power and commerce played a central role in this domination. Exported Etruscan objects of the period have been found in North Africa, Greece and the Aegean, Anatolia, Yugoslavia, France, and Spain; later they even reached the Black Sea. But land routes were well under control also, especially in the corridor leading through Rome and Latium down to Capua and the other Etruscanized cities of Campania. In northern Italy, Bologna (Felsina) was the principal city, and colonies such as the ones at nearby Marzabotto and at Adria and Spina on the Adriatic Sea represented significant posts along the northern trade network.

Almost from the beginning, the Etruscans must have been rivaled in their own seas by the Greeks, who, from the founding of Pithekoussai and Cumae, settled in numerous colonies in southern Italy, and by the Phoenicians, who had established Carthage about 800 BC. The Carthaginians claimed parts of Sicily, Corsica, and Sardinia as spheres of influence and dominated the seas west of these islands to Spain. The generally salutary trading relations among these three nations and the delicate balance of power were upset, however, in the Archaic period, as new waves of Greek colonists arrived. Phocaean Greeks established a colony on Corsica at Alalia (modern Aleria) which threatened both the Etruscans at Caere and the Carthaginians and led to a naval coalition between them. The ensuing battle in the seas off Corsica (*c.* 535 BC) had disastrous results for the Phocaeans, who emerged as victors but lost so many ships that they abandoned their colony and moved to southern Italy. The Carthaginians and Etruscans reasserted control over Corsica, and Etruscan might was to hold firm for another quarter of a century.

Organization. From the 6th century BC onward, territorial organization and political and economic initiative were concentrated in a limited number of large city-states in Etruria itself. These city-states, similar to the Greek poleis, consisted of an urban centre and a territory of fluctuating size. Numerous sources refer to a league of the "Twelve Peoples" of Etruria, formed for religious purposes but evidently having some political functions; it met annually at the chief sanctuary of the Etruscans, the Fanum Voltumnae, or shrine of Voltumna, near Volsinii. The precise location of the shrine is unknown, though it may have been in an area near modern Orvieto (believed by many to be the ancient Volsinii). As for the Twelve Peoples, no firm list of these has survived (indeed, they seem to have varied through the years), but they are likely to have come from the following major sites: Caere, Tarquinii, Vulci, Rusellae, Vetulonia, Populonia—all near the coast—and Veii, Volsinii, Clusium, Perusia (Perugia), Cortona, Arretium (Arezzo), Faesulae (Fiesole), and Volaterrae (Volterra)—all inland. There also are reports of corresponding Etruscan leagues in Campania and in northern Italy, but it is far more difficult to generate a list of Etruscan colonies or Etruscanized cities that would be likely candidates for these.

League of the Twelve Peoples

The names of some magistracies both in the league and in individual cities—such as *lauchme, zilath, maru,* and *purth*—are known, though there is little certainty as to their precise duties. *Lauchme* (Latin *lucumo*) was the Etruscan word for "king." The title of *zilath . . . rasnal,* translated into Latin as *praetor Etruriae* and meaning something like the "justice of Etruria," was evidently applied to the individual who presided over the league.

The men holding such magistracies belonged to the aristocracy, which derived its status from the continuity of the family. Onomastic formulas show that persons of free birth normally had two names. First came an individual name, or praenomen (relatively few of these are known: for men, Larth, Avle, Arnth, and Vel were frequent; for women, Larthia, Thanchvil, Ramtha, and Thana); it was followed by a family name, or nomen, derived from a personal name or perhaps the name of a god or a place. This system was in use by the second half of the 7th century, replacing the use of a single name (as in "Romulus" and "Remus") and reflecting the new complexity of relationships developing with urbanization. The Etruscans rarely used the cognomen (family nickname) employed by the Romans, but often inscriptions include the name of both the father (patronymic) and the mother (matronymic).

Etruscan women enjoyed an elevated status and a degree of liberation unknown to their counterparts in Rome and, especially, in Greece. They were allowed to own and openly display objects and clothing of a luxurious nature; they participated freely in public life, attending parties and theatrical performances; and—shocking to Greeks and Romans—they danced, drank, and rested in close physical contact with their husbands on the banqueting couches. Etruscan ladies were often literate, as one may deduce from the inscriptions on their mirrors, and even learned, if Livy's portrayal of Tanaquil as skilled in augury may be trusted. Their prominence in the family was a consistent feature of Etruscan aristocratic society and seems to have played a role in its stability and durability.

Etruscan women

Crisis and decline. The end of the 6th century and the beginning of the 5th was a turning point for Etruscan civilization. Several crises occurred at this time, from which the Etruscans never fully recovered and which in fact turned out to be only the first of numerous reverses they were to suffer in the ensuing centuries. The expulsion of the Tarquins from Rome (509 BC) deprived them of control over this strategic spot on the Tiber and also cut off their land route to Campania. Soon afterward, their naval supremacy also collapsed when the ships of the ambitious Hieron I of Syracuse inflicted a devastating loss on their fleet off Cumae in 474 BC. Completely out of touch with the Etruscan cities of Campania, they were unable to prevent a takeover of this area by restless Umbro-Sabellian tribes moving from the interior toward the coast.

All these reverses led to economic depression and a sharp interruption of trade for the cities on the coast and in the south and caused a redirection of commerce toward

the Adriatic harbour of Spina. The situation in the south deteriorated even further as Veii experienced periodic conflict with Rome, its close neighbour, and became the first Etruscan state to fall to this growing power in central Italy (396 BC).

Etruscan decline

A measure of prosperity had come to the Po valley and the Adriatic towns, but even this Etruscan vitality in the north was short-lived. The progressive infiltration and pressure of the Celts, who had penetrated and settled in the plain of the Po, eventually suffocated and overpowered the flourishing Etruscan urban communities, almost completely destroying their civilization by the mid-4th century BC and thus returning a large part of northern Italy to a protohistoric stage of culture. Meanwhile, the Gallic Senones firmly occupied the Picenum district on the Adriatic Sea, and Celtic incursions reached on the one hand Tyrrhenian Etruria and Rome (captured and burned about 390 BC) and on the other as far as Puglia.

In the 4th century BC ancient Italy had become profoundly transformed. The eastern Italic people of Umbro-Sabellian stock diffused over most of the peninsula; the Syracusan empire and lastly the growing power of Rome had replaced the Etruscans (and the Greek colonies of southern Italy) as the dominant force. The Etruscan world had been reduced to a circumscribed, regional sphere, secluded in its traditional values; this situation determined its progressive passage into the political system of Rome.

Within this context, Etruria experienced an economic recovery and a rebounding of the aristocracy. Tomb groups once again contain riches, and the sequence of painted tombs at Tarquinii, interrupted during the 5th century, resumes. All the same, there is a new atmosphere in these tombs; now one finds images of a grim afterlife, represented as an underworld replete with demons and overhung by dark clouds.

Renewed resistance to the power on the Tiber proved futile. Roman history is filled with records of victories and triumphs over Etruscan cities, especially in the south. Tarquinii sued for peace in 351 BC, and Caere was granted a truce in 353; there were triumphs over Rusellae in 302 and over Volaterrae in 298, with the final defeat of Rusellae coming in 294. Volsinii also was attacked in this year, and its fields devastated. During this same bleak period, Etruscan society was wracked with class struggles that eventually led to the development of a substantial freedman class, especially in northern Etruria, where numerous small rural settlements sprang up in the hills. In some cities, the aristocracy looked to Rome for assistance against the restless slave class. The noble Cilnii family at Arretium called for help with a revolt of the lower classes in 302 BC, while at Volsinii the situation deteriorated so badly that the Romans marched in and razed the city (265 BC), resettling its inhabitants in Volsinii Novi (probably Bolsena).

By the mid-3rd century all Etruria appears to have been pacified and firmly subjected to Roman hegemony. In most cases, the Etruscan cities and their territories preserved a formal autonomy as independent states with their own magistrates, thus passing an uneventful period in the 2nd century BC, when the sources are largely silent about Etruria.

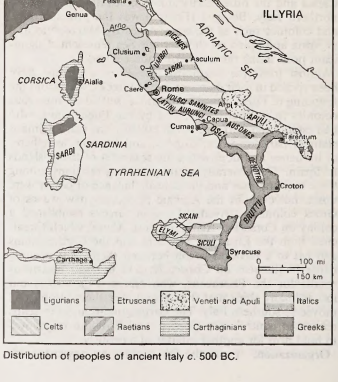
But the saddest chapter of all remained to be written in the 1st century BC. In 90 BC Rome granted citizenship to all Italic peoples, an act that in effect created total political unification of the Italic-Roman state and eliminated the last pretenses of autonomy in the Etruscan city-states. Northern Etruria, in addition, underwent a final devastation as it became the battleground for the opposing forces of the civil war of Marius and Sulla. Many Etruscan cities sided with Marius and were sacked and punished with all the vengeance the victorious Sulla could muster (80–79 BC). At Faesulae, Arretium, Volaterrae, and Clusium, the dictator confiscated and distributed territorial lands to soldiers from his 23 victorious legions. The new colonists brutally abused the old inhabitants and at the same time squandered their military rewards, sinking hopelessly into debt. Revolts and reprisals followed, but the agonizing process of Romanization was not actually completed until the reign of Augustus (31 BC–AD 14) brought new economic stability and reconciliation. By this time Latin had almost completely replaced the Etruscan language.

OTHER ITALIC PEOPLES

Local populations in areas colonized by Greece. The presence of the Siculi in Sicily and in the Italian peninsula is attested by the historical sources (Thucydides and Polybius). But the extent of their diffusion and their connections with other peoples of the peninsula (such as the Ligurians, the Itali, the Oenotrii, the Ausones) is more difficult to establish. A few small non-Greek inscriptions found in eastern Sicily and referable to the Siculi (the most noteworthy was found at Centuripe), coin legends, and Sicular words reported by Classical writers demonstrate the Indo-European character of the Sicular language, which seems to show an affinity with Latin and also has connections with the Umbro-Sabellian dialects. The immigration of the Siculi from the Italian peninsula into Sicily goes back to a prehistoric but not extremely early epoch; this assumption is based on the fact that there are some echoes of it in tradition and that a continental archaeological influence suddenly appears at the end of the Bronze Age. The characteristic Siculan iron culture, evident in the necropolises of Pantalica near Syracuse and of Finocchito near Noto, flourished between the 9th and the 5th centuries BC and was progressively submerged in the superior civilization of the Greeks. The Siculi

Evidence is quite scarce for the Siculi in the peninsula and for the other primitive indigenous populations of what is now Calabria and of Lucania (the Oenotrii, Ausones, Chones, Morgetes, and Itali) and Campania (the Ausones and Opici). Modern scholars have hypothesized that along the Tyrrhenian coastal arc there extended in earliest times a belt of paleo-Italic peoples (so-called western Italics or Proto-Latins) originally related to the Latins and distinct from the eastern Italic peoples inhabiting the Apennine and Adriatic regions. (The archaeological documentation consists of cemeteries with Iron Age graves, but there are also traces of cremation necropolises, especially in the province of Salerno and in Calabria.) Large fortified towns arose, particularly in the interior zones of Lucania. The ethnic individuality of these ancient peoples, however, was progressively obliterated between the 8th and the 4th century BC by the Greek penetration and by the expansion of the Etruscans and the eastern Italics (Samnites, Lucanians, Bruttii). Western Italic peoples

The Apulians. The inhabitants of the southeastern ex-



Distribution of peoples of ancient Italy *c.* 500 BC.

tremity of the Italian peninsula formed a definitely characterized group of populations that the ancients often called Iapyges (whence the geographic term Iapygia, in which "Apulia" may be recognized). The territory included the Salentini and the Messapii peoples in the Salentine Peninsula (ancient Calabia) and the Peucetii and the Dauni farther north. (Sometimes the designations Iapyges and Messapii are used with identical reference.) Ancient tradition insists upon an overseas origin for these tribes, held to be Cretan or Illyrian. The Iapygian, or more commonly, Messapian (Messapic) language is known from a considerable series of public funerary, votive, monetary, and other inscriptions written in the Greek alphabet and found in the Apulian area, especially in the Salentine Peninsula, from words reported by the ancient writers, and from toponomastic (local place-name) data. Messapian is without doubt an Indo-European language, distinct from Latin and from the Umbro-Sabellian dialects, with Balkan and central European analogies. This confirms the overseas provenance of the Iapyges from the Balkans, the more so because there existed in Illyria a tribe called the Iapodes and because a people known as the Iapuzkus lived farther north, on the Adriatic coast of Italy. Rather than a true immigration, however, there was a gradual prehistoric penetration of trans-Adriatic elements. The expansion of the Iapyges must have brought them to Lucania and even to what is now Calabria, as would be deduced from traditional and archaeological indications.

Apulian civilization

The Apulians' civilization, which was considerably influenced by that of the nearby Greek colonies, developed from the 9th to the 3rd centuries BC. In the most ancient period there were pit graves, sometimes in large stone tumuli. In the Siponto area, near what is now Manfredonia, the graves were accompanied by anthropomorphic stelae with geometric bas-reliefs. Geometrically painted ceramics in linear motifs persisted to the threshold of the Hellenistic age. Later graves took the form of large trunks and of catacombs with paintings on the sides. Burial was the disposition exclusively used.

Beginning in Archaic times, large cities developed, linked to each other by bonds of confederation. These included Herdonea (now Ordona), Canusium (Canosa), Rubi (Ruvo), Gnathia, Brundisium (Brindisi), Uria (Oria), Lupiae (Lecce), Rudiae, and Manduria. They preserved their independence, tenaciously defended against the Greeks, until the age of the Roman conquest.

The Latins. The Latin nation had a relatively limited territory, south of the Tiber, which was reduced, in historic times, by the invasion of the Volsci to the region between the Alban hills and the Aurunci mountains (the so-called Latium Novum). The principal Latin centres included Alba Longa, Tusculum, Lavinium, Ardea, Tibur (now Tivoli), and Praeneste (Palestrina) and the early Volscianized cities of Velitrae (Velletri), Signia (Segni), Cora (Cori), Satricum, Antium (Anzio), and Anxur (Terracina). The importance of the Latins is essentially linked with the fortunes of Rome, the forward bulwark of Latinity in the direction of the Etruscan realm. Intermixtures with the legends of the origin of Rome make the ethnographic traditions of Latium very diverse and complex. The linguistic evidence, which begins with inscriptions of the 7th to 6th century BC, indicates an individuality of the Latin world distinct from the neighbouring Etruscan and eastern Italic peoples.

The Latins had a federal organization, centred at the sanctuary of Jupiter on Albanus Mons. Their religious heritage survived in the beliefs and cults of the Roman world.

Latin culture

The most ancient Latin culture (9th–8th century BC) was characterized by cremation as the funeral rite, a practice it had in common with the cultures in the Etruscan and northern Italian territory, and by an iron culture showing affinities with the proto-Villanovan culture and with the cultures of Tyrrhenian southern Italy. Etruscan political control of Latium (probably 7th–6th century) coincided with an evident Etruscan cultural and artistic influence; while, from the south and from the sea, elements of Greek civilization penetrated, beginning with the alphabet.

North of Latium lived tribes ethnically akin to the Latins, with principal centres at Capena, Narce, and Falerii (whence the name Faliscans). Their political and cultural history merges with that of the Etruscans. The Faliscan dialect, known from inscriptions, was originally Latin but was contaminated and modified by eastern Italic and Etruscan elements.

The eastern Italics. A great part of the central and southern Italian peninsula was occupied in protohistoric and historic times by populations forming a vast ethnic and linguistic unit—the eastern Italics or Umbro-Sabellians. To the south, in the mountains of the Abruzzo, lived the Samnites, who later spread into Campania, Lucania, and what is now Calabria. In the centre were the Vestini, Paeligni, Marrucini, Marsi, Aequi, Volsci, and Sabini. Farther north lived the Umbri. The origin and relationship of all these peoples is unclear. Ancient ethnographic traditions bearing on central Italy link the Samnites to the Sabines and the Sabines to the Umbri, locating their primitive centre of dispersion in the Rieti basin and in the area of Amiternum. Their diffusion was attributed to the mass emigration of an entire generation in search of a new homeland (so-called sacred spring).

Origin of eastern Italics

The linguistic data prove the unity of the group of eastern-Italic idioms, belonging to the Indo-European stock but differing from the Latin. Within this group may be distinguished a southern variant (Sabellic or Oscan) known from an abundant harvest of epigraphic documents from Samnium, Campania, and southern Italy. This variant is to be attributed to the Samnites and to a large part of the minor stocks of central Italy, including the Sabines, known from isolated inscriptions in central Italy. On the other hand, a northern (Umbrian) variant is represented by inscriptions of Umbria—principally bronze tablets from Gubbio, inscribed between the 4th and 1st century BC by a brotherhood of Umbrian priests—and by a bronze tablet from Velletri. The eastern Italic words reported by the Classical writers, as well as toponomastics, confirm these conclusions.

Notwithstanding the original unity of stock and of language, these populations had diverse histories and cultures. The Samnites from Molise (the Caraceni, the Pentri, and the Frentani) in the 5th and 4th centuries BC occupied Campania—where they vanquished Etruscans and Greeks and assumed from the local tribes the name Opici, or Osci—as well as Lucania (with the Hirpini or Lucani), reaching what is now Calabria—where they took the name Bruttii—and finally Sicily. Defeated by Rome in the Samnite wars (4th and early 3rd centuries BC), the Samnites tried for the last time, in the period of the Social War (90–83 BC), to counterpose to the Romans an Italic nationality of their own. A considerable difference existed between the culture of the mountain Samnites—organized in confederate tribes centred on fortified villages and in the 5th and 4th centuries still retaining aspects of the "iron culture"—and the high civilization of the Campani and Lucani established in the ancient cities of Capua, Nola, Nocera, and Paestum and dominated by Greek and Etruscan influence.

Some central tribes—the Marrucini, Vestini, Paeligni and Marsi—appear to be linked historically, politically, and culturally to the Samnites. The case is different with the Sabines, the Aequi, and the Volsci, whose period of expansion (6th and 5th centuries) is closely connected with earliest Rome and who had early contact with the Etrusco-Latin civilization.

The Umbri

The diffusion of the Umbri toward the north and beyond the Apennines lent credence to the ancient traditions relating to the great size of their territory. The traditions, however, are more probably based on the fact that the name Umbri is derived from that of a most ancient population, probably not Indo-European and certainly not Italic, living in the Apennine region before the diffusion of the eastern Italics. The history of the Umbrian cities—Iguvium (now Gubbio), Hispellum (Spello), Spoletium (Spoleto), Tuder (Todi), and others—is known only beginning with the period of the struggle of the Etruscans and Gauls against Rome. Umbrian civilization is revealed by the Gubbio Tablets, a document unique in its kind. The Umbrian artistic and material culture derived in large part from that of Etruria.

The populations of the Picenum. In historic times, expansion of the eastern Italic peoples placed them firmly along the Adriatic coastal tract corresponding to what is now the Marches region. Epigraphic and archaeological data give evidence also of the presence in the Picenum (an ancient region between the Apennines and the Adriatic) of the trans-Adriatic Liburni and the pre-Indo-European Asili. It is possible that elements were established here that had come by sea from Illyria (the Gubbio tablets mention the Iapuzkus, whose name recalls that of the Illyrian Iapodes and of the Apulian Iapyges). Inscriptions in the southern Piceno, however, seem to exhibit a close kinship with the Umbro-Sabellic dialects.

Picenum civilization

A material civilization flowered between the 8th and 5th centuries in the northern Abruzzo and in the Marches. This civilization is represented by the rich funerary equipment of burial tombs, whose type and decoration present affinities with the iron culture of Tyrrhenian and northern Italy and with that of the Balkans and which show Greek influence. Cremation tombs of Villanovan type have been found at Fermo. Also noteworthy is the presence of stone funerary sculpture. North of Ancona is a cultural variant, particularly in the necropolis of Novilara near Pesaro, where inscriptions are in a dialect other than that of the southern Picenum and difficult to classify.

It may be held that the middle-Adriatic iron cultures expressed an early Archaic mixture of eastern Italic and trans-Adriatic peoples, influenced by the Etruscans and the Greeks. Contributing to their decline were the Gauls and Syracusans, who established themselves in this area in the 4th century BC. In the 3rd century, the Picenum was already totally conquered by the Romans.

The Veneti. Ancient tradition held the Veneti to be an Illyrian people who, coming from the east, took possession of the region named for them (Venetia). To them are linked the Histri, the Carni, and various Alpine tribes. (The name of the Veneti, or its root, is widely diffused in the ethnic onomastics of central Europe and even of Asia.) The Venetic language is known from funerary and votive inscriptions, from words cited by the Classical writers, and from onomastic and toponomastic data. It is an Indo-European language of Archaic type bearing similarities to the Latin and the Germanic.

The principal centres of the Veneti, located at the western margin of the territory, were Padua and Este. Their culture developed from the 9th century to the period of Romanization, with relationships with the Golasecca, Villanovan, and Etruscan cultures and with the transalpine Hallstatt culture. Maximum development occurred in the 6th–4th centuries BC; particularly noteworthy is the production of figured bronze *situlae* (conical vessels). In the final period, Gallic (Celtic) influences are found. Phenomena parallel to those of Este appear in Istria. The Veneti were horse breeders and peaceful traders and navigators; protected by the waters of the lower Po and the lower Adige, they preserved their independence against Etruscan expansion and Celtic invasion and in the 3rd century BC entered into peaceful alliance with Rome.

Veneti and Ligurians

The Ligurians. For the ancients, the name Ligures designated the peoples of northwestern Italy, including northern Tuscany, Liguria, Piedmont, and part of what is now Lombardy. Historical tradition also placed them in central Italy, while the Classical writers and toponomastic affinities give them a broader diffusion beyond the Alps. The Ligurians also included the peoples of Corsica. The more ancient Greeks gave all the peoples of the western world the common designation of Ligyes (*i.e.*, Ligurians).

Linguistic data—furnished by toponomastics, lexical survivals, and Ligurian words cited by Classical writers—betray the presence of a pre-Indo-European Mediterranean stratum akin to that of western Europe. Inscriptions found in upper Lombardy and in the Ticino exhibit Indo-European characteristics and in particular Celtic influences. Thus the Liguri seem to belong to an environment formed in northern Italy after the Celtic invasion and called Celto-Ligures.

The Etruscan expansion in the plain of the Po and the invasion of the Gauls confined the Ligurians between the Alps and the Apennines, where they offered such resistance to Roman penetration that they gained a reputation with the ancients for primitive fierceness. Among the more considerable Ligurian monuments are rock engravings and anthropomorphic sculptures analogous to those of southern France, found in Lunigiana and Corsica. Some of these artistic manifestations are repeated in territories farther east. But it remains doubtful whether the similar cultural imprint indicates an original identity of stock. Ligurian and Celto-Ligurian tombs of the Lombard lakes region, often holding cremations, reveal a special iron culture called the culture of Golasecca, while Ligurian sepulchres of the Italian Riviera and of Provence, also holding cremations, exhibit Etruscan and Celtic influences.

Ligurian culture

Populations of central northern Italy and of the Alps. The ethnography of the Po and Alpine regions is complex and obscure because of the early spread of Etruscan culture and colonization. The ancients record two major ethnic groups (aside from the Etruscans and the Veneti): the Euganei, inhabiting the plain and the Alpine foothills, and the Raeti, in the valleys of the Trentino and the Alto Adige. Minor peoples in the region belonged to one or the other of these stocks or to Ligurian stocks; with regard to many of these peoples, the sources speak of an Illyrian or an Etruscan origin.

Late inscriptions discovered in the Adige River valley and on the plain have a dialect showing some affinities to Etruscan. Some scholars see in this a blending of local and Etruscan elements, while others speak of an indigenous pre-Indo-European language with Indo-European influences. Primitive toponomastics confirm the existence of a linguistic stratum that could be defined as Raetian or Raeto-Euganean but distinguish it sharply from the Venetic and probably also from the Ligurian. Other inscriptions from the Val Camonica and the Garda district attest to a more noticeable Indo-European dialect, due perhaps to Celtic and Latin influences. To the west are the so-called Lepontian inscriptions.

Thus in the central Alpine and sub-Alpine area, there were original populations, different from the Veneti and the Etruscans, whose kinship with the Ligurians remains uncertain. The distinction between Euganean and Raetic tribes can be based only upon an approximate geographic criterion. To this original ethnic stratum may have belonged the most ancient inhabitants of the region, who settled there before the immigration of the Illyrian Veneti and the Etruscan conquest; certain cremation sepulchres of the Verona and Mantua regions may be attributed to them. Perhaps the existence of a Venetian goddess Reitia, recorded by Strabo and mentioned in inscriptions from Este, is some proof of a Raeto-Euganean cultural persistence in the territory occupied by the Veneti. (N.T.deG.)

Original central- and sub-Alpine populations

Ancient Rome

Rome must be considered one of the most successful imperial powers in history. In the course of centuries Rome grew from a small town on the Tiber River in central Italy into a vast empire that ultimately embraced England, all of continental Europe west of the Rhine and south of the Danube, most of Asia west of the Euphrates, northern Africa, and the islands of the Mediterranean. Unlike the Greeks, who excelled in intellectual and artistic endeavours, the Romans achieved greatness in their military, political, and social institutions. Roman society, during the republic, was governed by a strong military ethos. While this helps to explain the incessant warfare, it does not account for Rome's success as an imperial power. Unlike Greek city-states, which excluded foreigners and subjected peoples from political participation, Rome from its beginning incorporated conquered peoples into its social and political system. Allies and subjects who adopted Roman ways were eventually granted Roman citizenship. During the principate (see below), the seats in the Senate and even the imperial throne were occupied by persons from the Mediterranean realm outside Italy. The lasting effects of Roman rule in Europe can be seen in the geographic distribution of the Romance languages (Italian, French, Spanish, Portuguese, and Romanian), all of which evolved from Latin, the language of the Romans. The Western

alphabet of 26 letters and the calendar of 12 months and 365.25 days are only two simple examples of the cultural legacy which Rome has bequeathed Western civilization.

ROME FROM ITS ORIGINS TO 264 BC

Early Rome to 509 BC. *Early Italy.* When Italy emerged into the light of history about 700 BC, it was already inhabited by various peoples of different cultures and languages. Most natives of the country lived in villages or small towns, supported themselves by agriculture or animal husbandry (Italia means "Calf Land"), and spoke an Italic dialect belonging to the Indo-European family of languages. Oscan and Umbrian were closely related Italic dialects spoken by the inhabitants of the Apennines. The other two Italic dialects, Latin and Venetic, were likewise closely related to each other and were spoken, respectively, by the Latins of Latium (a plain of west-central Italy) and the people of northeastern Italy (near modern Venice). Iapyges and Messapii inhabited the southeastern coast. Their language resembled the speech of the Illyrians on the other side of the Adriatic. During the 5th century BC the Po valley of northern Italy (Cisalpine Gaul) was occupied by Gallic tribes who spoke Celtic and who had migrated across the Alps from continental Europe. As noted earlier, the Etruscans were the first highly civilized people of Italy and were the only inhabitants who did not speak an Indo-European language. By 700 BC several Greek colonies were established along the southern coast. Both Greeks and Phoenicians were actively engaged in trade with the Italian natives.

Native Italian peoples

Modern historical analysis is making rapid progress in showing how Rome's early development occurred in a multicultural environment and was particularly influenced by the higher civilizations of the Etruscans to the north and the Greeks to the south. Roman religion was indebted to the beliefs and practices of the Etruscans. The Romans borrowed and adapted the alphabet from the Etruscans, who in turn had borrowed and adapted it from the Greek colonies of Italy. Senior officials of the Roman Republic derived their insignia from the Etruscans: curule chair, purple-bordered toga (*toga praetexta*), and bundle of rods (*fasces*). Gladiatorial combats and the military triumph (see below) were other customs adopted from the Etruscans. Rome lay 12 miles inland from the sea on the Tiber River, the border between Latium and Etruria. Because the site commanded a convenient river crossing and lay on a land route from the Apennines to the sea, it formed the meeting point of three distinct peoples: Latins, Etruscans, and Sabines. Though Latin in speech and culture, the Roman population must have been somewhat diverse from earliest times, a circumstance that may help to account for the openness of Roman society in historical times.

Historical sources on early Rome. The regal period (753–509 BC) and the early republic (509–280 BC) are the most poorly documented periods of Roman history because historical accounts of Rome were not written until much later. Greek historians did not take serious notice of Rome until the Pyrrhic War (280–275 BC), when Rome was completing its conquest of Italy and was fighting against the Greek city of Tarentum in southern Italy. Rome's first native historian, a senator named Quintus Fabius Pictor, lived and wrote even later, during the Second Punic War (218–201 BC). Thus historical writing at Rome did not begin until after Rome had completed its conquest of Italy, had emerged as a major power of the ancient world, and was engaged in a titanic struggle with Carthage for control of the western Mediterranean. Fabius Pictor's history, which began with the city's mythical Trojan ancestry and narrated events up to his own day, established the form of subsequent histories of Rome. During the last 200 years BC, 16 other Romans wrote similarly inclusive narratives. All these works are now collectively termed "the Roman annalistic tradition" because many of them attempted to give a year-by-year (or annalistic) account of Roman affairs for the republic.

The Roman annalistic tradition

Although none of these histories are fully preserved, the first 10 books of Livy, one of Rome's greatest historians, are extant and cover Roman affairs from earliest times down to the year 293 BC (extant are also Books 21 to 45 treating the events from 218 BC to 167 BC). Since Livy wrote during the reign of the emperor Augustus (27 BC–AD 14), he was separated by 200 years from Fabius Pictor, who, in turn, had lived long after many of the events his history described. Thus, in writing about early Rome, ancient historians were confronted with great difficulties in ascertaining the truth. They possessed a list of annual magistrates from the beginning of the republic onward (the consular *fasti*), which formed the chronological framework of their accounts. Religious records and the texts of some laws and treaties provided a bare outline of major events. Ancient historians fleshed out this meagre factual material with both native and Greek folklore. Consequently, over time, historical facts about early Rome often suffered from patriotic or face-saving reinterpretations involving exaggeration of the truth, suppression of embarrassing facts, and invention.

The evidence for the annalistic tradition shows that the Roman histories written during the 2nd century BC were relatively brief resumes of facts and stories. Yet in the course of the 1st century BC Roman writers were increasingly influenced by Greek rhetorical training, with the result that their histories became greatly expanded in length; included in them were fictitious speeches and lengthy narratives of spurious battles and political confrontations, which, however, reflect the military and political conditions and controversies of the late republic rather than accurately portraying the events of early Rome. Livy's history of early Rome, for example, is a blend of some facts and much fiction. Since it is often difficult to separate fact from fiction in his works and doing so involves personal judgment, modern scholars have disagreed about many aspects of early Roman history and will continue to do so.

Rome's foundation myth. Although Greek historians did not write seriously about Rome until the Pyrrhic War, they were aware of Rome's existence long before then. In accordance with their custom of explaining the origin of the foreign peoples they encountered by connecting them with the wanderings of one of their own mythical heroes, such as Jason and the Argonauts, Heracles, or Odysseus, Greek writers from the 5th century BC onward invented at least 25 different myths to account for Rome's foundation. In one of the earliest accounts (Hellanicus of Lesbos), which became accepted, the Trojan hero Aeneas and some followers escaped the Greek destruction of Troy; after wandering about the Mediterranean for some years, they settled in central Italy, where they intermarried with the native population and became the Latins.

Although the connection between Rome and Troy is unhistorical, the Romans of later time were so flattered by this illustrious mythical pedigree that they readily accepted it and incorporated it into their own folklore about the beginning of their city. Roman historians knew that the republic had begun about 500 BC, because their annual list of magistrates went back that far. Before that time, they thought, Rome had been ruled by seven kings in succession. By using Greek methods of genealogical reckoning, they estimated that seven kings would have ruled about 250 years, thus making Rome's regal period begin in the middle of the 8th century BC. Ancient historians initially differed concerning the precise date of Rome's foundation, ranging from as early as 814 BC (Timaeus) to as late as 728 BC (Cincius Alimentus). By the end of the republic, it was generally accepted that Rome had been founded in 753 BC and that the republic had begun in 509 BC.

Dating efforts by ancient historians

Since the generally accepted date of Troy's destruction was 1184 BC, Roman historians maintained Troy's unhistorical connection with Rome by inventing a series of fictitious kings who were supposed to have descended from the Trojan Aeneas and ruled the Latin town of Alba Longa for the intervening 431 years (1184–753 BC) until the last of the royal line, the twin brothers Romulus and Remus, founded their own city, Rome, on the Palatine Hill. According to tradition, the twins, believed to have been the children of the god Mars, were set adrift in a basket on the Tiber by the king of Alba; they survived, however, being nursed by a she-wolf, and lived to overthrow the wicked king. In the course of founding Rome the brothers quar-

reled, and Romulus slew Remus. This story was a Roman adaptation of a widespread ancient Mediterranean folktale told of many national leaders, such as the Akkadian king Sargon (*c.* 2300 BC), the biblical Moses, the Persian king Cyrus the Great, the Theban king Oedipus, and the twins Neleus and Pelias of Greek mythology.

The regal period, 753–509 BC. Romulus, Rome's first king according to tradition, was the invention of later ancient historians. His name, which is not even proper Latin, was designed to explain the origin of Rome's name. His fictitious reign was filled with deeds expected of an ancient city founder and the son of a war god. Thus he was described as having established Rome's early political, military, and social institutions and as having waged war against neighbouring states. Romulus was also thought to have shared his royal power for a time with a Sabine named Titus Tatius. The name may be that of an authentic ruler of early Rome, perhaps Rome's first real king; nothing, however, was known about him in later centuries, and his reign was therefore lumped together with that of Romulus.

The names of the other six kings are authentic and were remembered by the Romans, but few reliable details were known about their reigns. However, since the later Romans wished to have explanations for their early customs and institutions, historians ascribed various innovations to these kings, often in stereotypical and erroneous ways. The three kings after Romulus are still hardly more than names, but the recorded deeds of the last three kings are more historical and can, to some extent, be checked by archaeological evidence.

According to ancient tradition, the warlike founder Romulus was succeeded by the Sabine Numa Pompilius, whose reign was characterized by complete tranquility and peace. Numa was supposed to have created virtually all of Rome's religious institutions and practices. The tradition of his religiosity probably derives from the erroneous connection by the ancients of his name with the Latin word *numen,* meaning divine power. Numa was succeeded by Tullus Hostilius, whose reign was filled with warlike exploits, probably because the name Hostilius was later interpreted to suggest hostility and belligerence. Tullus was followed by Ancus Marcius, who was believed to have been the grandson of Numa. His reign combined the characteristics of those of his two predecessors—namely religious innovations as well as warfare.

Archaeological evidence for early Rome is scattered and limited because it has proven difficult to conduct extensive excavations at sites still occupied by later buildings. What evidence exists is often ambiguous and cannot be correlated easily to the ancient literary tradition; it can, however, sometimes confirm or contradict aspects of the ancient historical account. For example, it confirms that the earliest settlement was a simple village of thatched huts on the Palatine Hill (one of the seven hills eventually occupied by the city of Rome), but it dates the beginning of the village to the 10th or 9th century BC, not the mid-8th century. Rome therefore cannot have been ruled by a succession of only seven kings down to the end of the 6th century BC. Archaeology also shows that the Esquiline Hill was next inhabited, thus disproving the ancient account which maintained that the Quirinal Hill was settled after the Palatine. Around 670–660 BC the Palatine settlement expanded down into the valley of the later Forum Romanum and became a town of artisans living in houses with stone foundations. The material culture testifies to the existence of some trade as well as to Etruscan and Greek influence. Archaeology of other Latin sites suggests that Rome at this time was a typical Latin community. In another major transition spanning the 6th century the Latin town was gradually transformed into a real city. The swampy Forum valley was drained and paved to become the city's public centre. There are clear signs of major temple construction. Pottery and architectural remains indicate vigorous trade with the Greeks and Etruscans, as well as local work done under their influence.

Archaeological evidence

Rome's urban transformation was carried out by its last three kings: Lucius Tarquinius Priscus (Tarquin the Elder), Servius Tullius, and Lucius Tarquinius Superbus (Tarquin the Proud). According to ancient tradition, the two Tarquins were father and son and came from Etruria. One tradition made Servius Tullius a Latin; another described him as an Etruscan named Mastarna. All three kings were supposed to have been great city planners and organizers (a tradition that has been confirmed by archaeology). Their Etruscan origin is rendered plausible by Rome's proximity to Etruria, Rome's growing geographic significance, and the public works that were carried out by the kings themselves. The latter were characteristic of contemporary Etruscan cities. It would thus appear that during the 6th century BC some Etruscan adventurers took over the site of Rome and transformed it into a city along Etruscan lines.



A funerary urn in the shape of one of the earliest types of Roman dwelling, *c.* 8th century BC. In the Museum of Antiquities from the Palatine, Rome.

Courtesy Sopraintendenza Archeologica di Roma

Early centuries of the Roman Republic. *Foundation of the republic.* The ancient historians depicted Rome's first six kings as benevolent and just rulers but the last one as a cruel tyrant who murdered his predecessor Servius Tullius, usurped the kingship, terrorized the Senate, and oppressed the common people with public works. He supposedly was overthrown by a popular uprising ignited by the rape of a virtuous noblewoman, Lucretia, by the king's son. The reign of Tarquinius Superbus was described in the stereotypical terms of a Greek tyranny in order to explain the major political transition from the monarchy to the republic in accordance with Greek political theory concerning constitutional evolution from monarchy to tyranny to aristocracy. This explanation provided later Romans with a satisfying patriotic story of despotism giving way to liberty; it is probably unhistorical, however, and merely a Roman adaptation of a well-known Greek story of a love affair in Athens that led to the murder of the tyrant's brother and the tyrant's eventual downfall. According to ancient tradition, as soon as the Romans had expelled their last tyrannical king, the king of the Etruscan city of Clusium, Lars Porsenna, attacked and besieged Rome. The city was gallantly defended by Horatius Cocles, who sacrificed his life in defense of the bridge across the Tiber, and Mucius Scaevola, who attempted to assassinate Porsenna in his own camp. When arrested before accomplishing the deed, he demonstrated his courage by voluntarily burning off his right hand in a nearby fire. As a result of such Roman heroism, Porsenna was supposed to have made peace with Rome and withdrawn his army.

One prevalent modern view is that the monarchy at Rome was incidentally terminated through military defeat and foreign intervention. This theory sees Rome as a site highly prized by the Etruscans of the 6th century BC, who are known to have extended their power and influence at the time across the Tiber into Latium and even farther south into Campania. Toward the end of the 6th century, Rome may have been involved in a war against King Porsenna of Clusium, who defeated the Romans, seized the city, and expelled its last king. Before Porsenna could establish himself as monarch, he was forced to withdraw, leaving Rome kingless. In fact, Porsenna is known to have suffered a serious defeat at the hands of the combined

Termination of the monarchy

forces of the other Latins and the Greeks of Campanian Cumae. Rather than restoring Tarquin from exile to power, the Romans replaced the kingship with two annually elected magistrates called consuls.

The struggle of the orders. As the Roman state grew in size and power during the early republic (509–280 BC), new offices and institutions were created, and old ones were adapted to cope with the changing military, political, social, and economic needs of the state and its populace. According to the annalistic tradition, all these changes and innovations resulted from a political struggle between two social orders, the patricians and the plebeians, that is thought to have begun during the first years of the republic and lasted for more than 200 years. In the beginning, the patricians were supposed to have enjoyed a monopoly of power (the consulship, the Senate, and all religious offices), whereas the plebeians began with nothing except the right to vote in the assemblies. During the course of the struggle the plebeians, however, were believed to have won concessions gradually from the patricians through political agitation and confrontation, and they eventually attained legal equality with them. Thus ancient historians, such as Livy, explained all aspects of early Rome's internal political development in terms of a single sustained social movement.

As tradition has it, the distinction between patricians and plebeians was as old as Rome itself and had been instituted by Romulus. The actual historical dating and explanation of this distinction still constitutes the single biggest unsolved problem of early Roman history. The distinction existed during the middle and late republic, but modern scholars do not agree on when or how it arose; they are increasingly inclined to think that it originated and evolved slowly during the early republic. By the time of the middle and late republic, it was largely meaningless. At that point only about one dozen Roman families were patrician, all others being plebeian. Both patrician and plebeian families made up the nobility, which consisted simply of all descendants of consuls. The term "patrician," therefore, was not synonymous with "noble" and should not be confused with it: the patricians formed only a part of the Roman nobility of the middle and late republic. The only difference between patricians and plebeians in later times was that each group was either entitled to or debarred from holding certain minor offices.

The discrepancies, inconsistencies, and logical fallacies in Livy's account of the early republic make it evident that the annalistic tradition's thesis of a struggle of the orders is a gross oversimplification of a highly complex series of events that had no single cause. Tensions certainly existed; no state can experience 200 years of history without some degree of social conflict and economic unrest. In fact, legal sources indicate that the law of debt in early Rome was extremely harsh and must have sometimes created much hardship. Yet it is impossible to believe that all aspects of early Rome's internal political development resulted from one cause. Early documents, if available, would have told the later annalistic historians little more than that a certain office had been created or some law passed. An explanation of causality could have been supplied only by folklore or by the imagination of the historian himself, neither of which can be relied upon. Livy's descriptions of early republican political crises evince the political rhetoric and tactics of the late republic and therefore cannot be given credence without justification. For example, early republican agrarian legislation is narrated in late republican terms. Early republican conflicts between plebeian tribunes and the Senate are likewise patterned after the politics of the Optimates and Populares of the late republic. Caution therefore must be exercised in examining early Rome's internal development. Many of the major innovations recorded in the ancient tradition can be accepted, but the ancient interpretation of these facts cannot go unchallenged.

The consulship. The later Romans viewed the abolition of the kingship and its replacement by the consulship as marking the beginning of the republic. The king's religious functions were henceforth performed by a priest-king (*rex sacrorum*), who held office for life. The king's military power (*imperium*) was bestowed upon two annually elected magistrates called consuls. They were always regarded as the chief magistrates of the republic, so much so that the names of each pair were given to their year of office for purposes of dating. Thus careful records were kept of these names, which later formed the chronological basis for ancient histories of the republic. The consuls were primarily generals who led Rome's armies in war. They were therefore elected by the centuriate assembly—that is, the Roman army organized into a voting body. The two consuls possessed equal power. Such collegiality was basic to almost all Roman public offices; it served to check abuses of power because one magistrate's actions could be obstructed by his colleague.

According to the annalistic tradition, the first plebeian consul was elected for 366 BC. All consuls before that time were thought to have been patrician, and one major aspect of the struggle of the orders was supposed to have been the plebeians' persistent agitation to make the office open to them. However, if the classification of patrician and plebeian names known for the middle and late republic is applied to the consular list for the years 509–445 BC, plebeian names are well represented (30 percent). It is likely that there never was a prohibition against plebeians holding the consulship. The distinction between patrician and plebeian families may have become fixed only by the middle of the 4th century BC; and the law of that time (367 BC), which specified that one of the consuls was to be plebeian, may have done nothing more than to guarantee legally that both groups of the nobility would have an equal share in the state's highest office.

The dictatorship. Despite the advantages of consular collegiality, in military emergencies unity of command was sometimes necessary. Rome's solution to this problem was the appointment of a dictator in place of the consuls. According to ancient tradition, the office of dictator was created in 501 BC, and it was used periodically down to the Second Punic War. The dictator held supreme military command for no longer than six months. He was also termed the master of the army (*magister populi*), and he appointed a subordinate cavalry commander, the master of horse (*magister equitum*). The office was thoroughly constitutional and should not be confused with the late republican dictatorships of Sulla and Caesar, which were simply legalizations of autocratic power obtained through military usurpation.

Rule during military emergencies

The Senate. The Senate may have existed under the monarchy and served as an advisory council for the king. Its name suggests that it was originally composed of elderly men (*senes*), whose age and knowledge of traditions must have been highly valued in a preliterate society. During the republic, the Senate was composed of members from the leading families. Its size during the early republic is unknown. Ancient sources indicate that it numbered about 300 during the middle republic. Its members were collectively termed *patres et conscripti* ("the fathers and the enrolled"), suggesting that the Senate was initially composed of two different groups. Since the term "patrician" was derived from *patres* and seems to have originally meant "a member of the *patres*," the dichotomy probably somehow involved the distinction between patricians and plebeians. During the republic the Senate advised both magistrates and the Roman people. Although in theory the people were sovereign (see below) and the Senate only offered advice, in actual practice the Senate wielded enormous power because of the collective prestige of its members. It was by far the most important deliberative body in the Roman state, summoned into session by a magistrate who submitted matters to it for discussion and debate. Whatever a majority voted in favour of was termed "the Senate's advice" (*senatus consultum*). These advisory decrees were directed to a magistrate or the Roman people. In most instances, they were either implemented by a magistrate or submitted by him to the people for enactment into law.

The popular assemblies. During the republic two different assemblies elected magistrates, exercised legislative power, and made other important decisions. Only adult male Roman citizens could attend the assemblies in Rome and exercise the right to vote. The assemblies were or-

ganized according to the principle of the group vote. Although each person cast one vote, he did so within a larger voting unit. The majority vote of the unit became its vote, and a majority of unit votes was needed to decide an issue.

The centuriate assembly (*comitia centuriata*), as stated, was military in nature and composed of voting groups called centuries (military units). Because of its military character, it always met outside the sacred boundary of the city (*pomerium*) in the Field of Mars (Campus Martius). It voted on war and peace and elected all magistrates who exercised imperium (consuls, praetors, censors, and curule aediles). Before the creation of criminal courts during the late republic, it sat as a high court and exercised capital jurisdiction. Although it could legislate, this function was usually performed by the tribal assembly.

The centuriate assembly

The centuriate assembly evolved through different stages during the early republic, but information exists only about its final organization. It may have begun as the citizen army meeting under arms to elect its commander and to decide on war or peace. During historical times the assembly had a complex organization. All voting citizens were placed into one of five economic classes according to wealth. Each class was allotted varying numbers of centuries, and the entire assembly consisted of 193 units. The first (and richest) class of citizens was distributed among 80 centuries; the second, third, and fourth classes were each assigned 20 units. The fifth class, comprising the poorest persons in the army, was allotted 30 centuries. In addition, there were 18 centuries of knights—men wealthy enough to afford a horse for cavalry service—and five other centuries, one of which comprised the *proletarii*, or landless people too poor to serve in the army. The knights voted together with the first class, and voting proceeded from richest to poorest. Because the knights and the first class controlled 98 units, they were the dominant group in the assembly, though they constituted the smallest portion of the citizen body. The assembly was deliberately designed to give the greater authority to the wealthier element and was responsible for maintaining the political supremacy of the established nobility.

The tribal assembly (*comitia tributa*) was a nonmilitary civilian assembly. It accordingly met within the city inside the pomerium and elected magistrates who did not exercise imperium (plebeian tribunes, plebeian aediles, and quaestors). It did most of the legislating and sat as a court for serious public offenses involving monetary fines.

The tribal assembly was more democratic in its organization than the centuriate assembly. The territory of the Roman state was divided into geographic districts called tribes, and people voted in these units according to residence. The city was divided into four urban tribes. During the 5th century BC, the surrounding countryside formed 17 rustic tribes. With the expansion of Roman territory in central Italy (387–241 BC), 14 rustic tribes were added, thus gradually increasing the assembly to 35 units, a number never exceeded.

The plebeian tribunate. According to the annalistic tradition, one of the most important events in the struggle of the orders was the creation of the plebeian tribunate. After being worn down by military service, bad economic conditions, and the rigours of early Rome's debt law, the plebeians in 494 BC seceded in a body from the city to the Sacred Mount, located three miles from Rome. There they pitched camp and elected their own officials for their future protection. Because the state was threatened with an enemy attack, the Senate was forced to allow the plebeians to have their own officials, the tribunes of the plebs.

Tribunes of the plebs

Initially there were only 2 tribunes of the plebs, but their number increased to 5 in 471 BC and to 10 in 457 BC. They had no insignia of office, like the consuls, but they were regarded as sacrosanct. Whoever physically harmed them could be killed with impunity. They had the right to intercede on a citizen's behalf against the action of a consul, but their powers were valid only within one mile from the pomerium. They convoked the tribal assembly and submitted bills to it for legislation. Tribunes prosecuted other magistrates before the assembled people for misconduct in office. They could also veto the action of another tribune (*veto* meaning "I forbid"). Two plebeian aediles served as their assistants in managing the affairs of the city. Although they were thought of as the champions of the people, persons elected to this office came from aristocratic families and generally favoured the status quo. Nevertheless, the office could be and sometimes was used by young aspiring aristocrats to make a name for themselves by taking up populist causes in opposition to the nobility.

Modern scholars disagree about the authenticity of the annalistic account concerning the plebs' first secession and the creation of the plebeian tribunate. The tradition presented this as the first of three secessions, the other two allegedly occurring in 449 and 287 BC. The second secession is clearly fictitious. Many scholars regard the first one as a later annalistic invention as well, accepting only the last one as historical. Although the first secession is explained in terms resembling the conditions of the later Gracchan agrarian crisis (see below), given the harshness of early Roman debt laws and food shortages recorded by the sources for 492 and 488 BC (information likely to be preserved in contemporary religious records), social and economic unrest could have contributed to the creation of the office. However, the urban-civilian character of the plebeian tribunate complements the extra-urban military nature of the consulship so nicely that the two offices may have originally been designed to function cooperatively to satisfy the needs of the state rather than to be antagonistic to one another.

The Law of the Twelve Tables. The next major episode after the creation of the plebeian tribunate in the annalistic version of the struggle of the orders involved the first systematic codification of Roman law. The plebeians were supposed to have desired a written law code in which consular imperium would be circumscribed to guard against abuses. After years of tribunician agitation the Senate finally agreed. A special board of 10 men (*decemviri*) was appointed for 451 BC to draw up a law code. Since their task was not done after one year, a second board of 10 was appointed to finish the job, but they became tyrannical and stayed in office beyond their time. They were finally forced out of power when one commissioner's cruel lust for an innocent maiden named Verginia so outraged the people that they seceded for a second time.

Provisions of the law code

The law code was inscribed upon 12 bronze tablets and publicly displayed in the Forum. Its provisions concerned legal procedure, debt foreclosure, paternal authority over children, property rights, inheritance, funerary regulations, and various major and minor offenses. Although many of its provisions became outmoded and were modified or replaced in later times, the Law of the Twelve Tables formed the basis of all subsequent Roman private law.

Because the law code seems not to have had any specific provisions concerning consular imperium, the annalistic explanation for the codification appears suspect. The story of the second tyrannical board of 10 is an annalistic invention patterned after the 30 tyrants of Athenian history. The tale of Verginia is likewise modeled after the story of Lucretia and the overthrow of Rome's last king. Thus the second secession, which is an integral part of the story, cannot be regarded as historical. On the basis of existing evidence, one cannot say whether the law code resulted from any social or economic causes. Rome was a growing city and may simply have been in need of a systematic body of law.

Military tribunes with consular power. The creation of the office of military tribunes with consular power in 445 BC was believed to have involved the struggle of the orders. The annalistic tradition portrayed the innovation as resulting from a political compromise between plebeian tribunes, demanding access to the consulship, and the Senate, trying to maintain the patrician monopoly of the office. Henceforth, each year the people were to decide whether to elect two patrician consuls or military tribunes with consular power who could be patricians or plebeians. The list of magistrates for 444 to 367 BC shows that the chief magistracy alternated between consuls and military tribunes. Consuls were more frequently elected down to 426 but rarely thereafter. At first there were three military

tribunes, but the number increased to four in 426, and to six in 406. The consular tribunate was abolished in 367 BC and replaced by the consulship.

Livy indicates that according to some sources the consular tribunate was created because Rome was faced with three wars simultaneously. Because there is evidence that there was no prohibition against plebeians becoming consuls, scholars have suggested that the reason for the innovation was the growing military and administrative needs of the Roman state; this view is corroborated by other data. Beginning in 447 BC two quaestors were elected as financial officials of the consuls, and the number increased to four in 421 BC. Beginning in 443 BC two censors were elected about every five years and held office for 18 months. They drew up official lists of Roman citizens, assessed the value of their property, and assigned them to their proper tribe and century within the tribal and centuriate assemblies. The increase in the number of military tribunes coincided with Rome's first two major wars, against Fidenae and Veii. In 366 BC six undifferentiated military tribunes were replaced with five magistrates that had specific functions: two consuls for conducting wars, an urban praetor who handled lawsuits in Rome, and two curule aediles who managed various affairs in the city. In 362 BC the Romans began to elect annually six military tribunes as subordinate officers of the consuls.

Social and economic changes. The law reinstating the consulship was one of three tribunician bills, the so-called Licinio-Sextian Rogations of 367 BC. Another forbade citizens to rent more than 500 *iugera* (330 acres) of public land, and the third provided for the alleviation of indebtedness. The historicity of the second bill has often been questioned, but the great increase in the size of Roman territory resulting from Rome's conquest of Veii renders this law plausible. The law concerning indebtedness is probably historical as well, since other data suggest that debt was a problem in mid-4th-century Rome. In 352 BC a five-man commission was appointed to extend public credit in order to reduce private indebtedness. A Genucian law of 342 BC (named after Genucius, tribune of that year) temporarily suspended the charging of interest on loans. In 326 or 313 BC a Poetelian law ameliorated the harsh conditions of the Twelve Tables regarding debt servitude by outlawing the use of chains to confine debt bondsmen.

Rome's economic advancement is reflected in its replacement of a cumbersome bronze currency with silver coinage adopted from the Greek states of southern Italy, the so-called Romano-Campanian didrachms. The date of this innovation is disputed. Modern estimates range from the First Samnite War to the Pyrrhic War. Rome was no longer a small town of central Italy but rather was quickly becoming the master of the Italian peninsula and was taking its place in the larger Mediterranean world.

Romano-Campanian didrachms

The process of expansion is well illustrated by innovations in Roman private law about 300 BC. Since legal business could be conducted only on certain days (*dies fasti*), knowledge of the calendar was important for litigation. In early times the *rex sacrorum* at the beginning of each month orally proclaimed in Rome before the assembled people the official calendar for that month. Though suited for a small agricultural community, this parochial procedure became increasingly unsuitable as Roman territory grew and more citizens lived farther from Rome. In 304 BC a curule aedile named Gnaeus Flavius upset conservative opinion but performed a great public service by erecting an inscription of the calendar in the Roman Forum for permanent display. From early times, Roman private law and legal procedure had largely been controlled and developed by the priesthood of pontiffs. In 300 BC the Ogulnian law (after the tribunes Gnaeus and Quintus Ogulnius) ended the patrician monopoly of two priestly colleges by increasing the number of pontiffs from four to eight and the number of augurs from four to nine and by specifying that the new priests were to be plebeian.

In 287 BC the third (and perhaps the only historical) secession of the plebs occurred. Since Livy's account has not survived, detailed knowledge about this event is lacking. One source suggests that debt caused the secession. Many sources state that the crisis was ended by the passage of the Hortensian law (after Quintus Hortensius, dictator for 287), which was thought to have given enactments of the tribal assembly the same force as resolutions of the centuriate assembly. However, since similar measures were supposed to have been enacted in 449 and 339 BC, doubt persists about the meaning of these laws. It is possible that no difference ever existed in the degree of legal authority of the two assemblies. The three laws could be annalistic misinterpretations of a provision of the Twelve Tables specifying that what the people decided last should be binding. One source indicates that the Hortensian law made all assembly days eligible for legal business. If debt played a role in the secession, the Hortensian law may have been designed to reduce the backlog of lawsuits in the praetor's court in Rome.

The Latin League. Although the Latins dwelled in politically independent towns, their common language and culture produced cooperation in religion, law, and warfare. All Latins could participate in the cults of commonly worshiped divinities, such as the cult of the Penates of Lavinium, Juno of Lanuvium, and Diana (celebrated at both Aricia and Rome). Latins freely intermarried without legal complications. When visiting another Latin town, they could buy, sell, litigate, and even vote with equal freedom. If a Latin took up permanent residence in another Latin community, he became a full citizen of his new home. Although the Latin states occasionally waged war among themselves, in times of common danger they banded together for mutual defense. Each state contributed military forces according to its strength. The command of all forces was entrusted by common assent to a single person from one of the Latin towns. Sometimes the Latins even founded colonies upon hostile territory as military outposts, which became new, independent Latin states, enjoying the same rights as all the other ones. Modern scholars use the term "Latin League" to describe this collection of rights and duties.

Rights and duties of the Latins

According to ancient tradition, Rome's last three kings not only transformed Rome into a real city but also made it the leader of the Latin League. There is probably exaggeration in this claim. Roman historians were eager to portray early Rome as destined for future greatness and as more powerful than it actually was. Rome certainly became one of the more important states in Latium during the 6th century, but Tibur, Praeneste, and Tusculum were equally important and long remained so. By the terms of the first treaty between Rome and Carthage (509 BC), recorded by the Greek historian Polybius (*c.* 150 BC), the Romans (or perhaps more accurately, the Latins generally) claimed a coastal strip 70 miles south of the Tiber River as their sphere of influence not to be encroached upon by the Carthaginians.

Rome's rapid rise during the 6th century was the achievement of its Etruscan overlords, and the city quickly declined with the collapse of Etruscan power in Campania and Latium about 500 BC. Immediately after the fall of the Roman monarchy, amid Porsenna's conquest of Rome, his defeat by the Latins, and his subsequent withdrawal, the plain of Latium began to be threatened by surrounding hill tribes (Sabines, Aequi, and Volsci), who experienced overpopulation and tried to acquire more land. Thus Rome's external affairs during the 5th century largely revolved around its military assistance to the Latin League to hold back these invaders. Many details in Livy's account of this fighting are, however, unreliable. In order to have a literary theme worthy of Rome's later greatness, Livy's annalistic sources had described these conflicts in the most grandiose terms. Yet the armies, military ranks, castrametation (*i.e.*, techniques in making and fortifying encampments), and tactics described belong to the late republic, not the Rome of the 5th century.

Roman expansion in Italy. Toward the end of the 5th century, while Rome and the Latins were still defending themselves against the Volsci and the Aequi, the Romans began to expand at the expense of Etruscan states. Rome's incessant warfare and expansion during the republic has spawned modern debate about the nature of Roman imperialism. Ancient Roman historians, who were often patriotic senators, believed that Rome always waged just

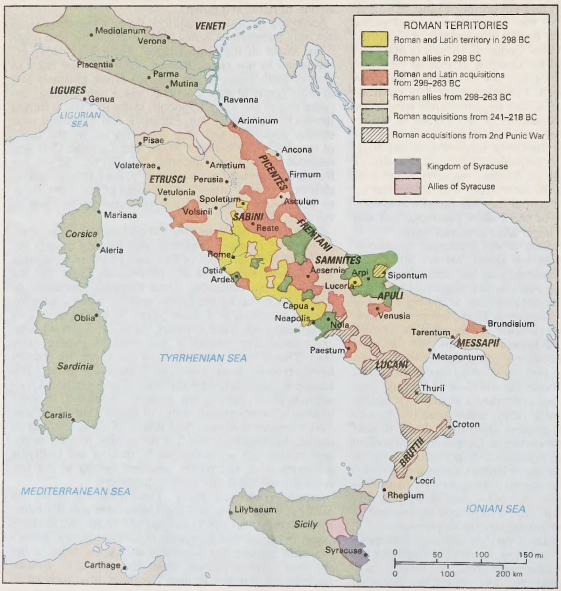
wars in self-defense, and they wrote their accounts accordingly, distorting or suppressing facts that did not fit this view. The modern thesis of Roman defensive imperialism, which followed this ancient bias, is now largely discredited. Only the fighting in the 5th century BC and the later wars against the Gauls can clearly be so characterized. Rome's relentless expansion was more often responsible for provoking its neighbours to fight in self-defense. Roman consuls, who led the legions into battle, often advocated war because victory gained them personal glory. Members of the centuriate assembly, which, as noted above, decided war and peace, may sometimes have voted for war in expectation that it would lead to personal enrichment through seizure and distribution of booty. The evidence concerning Roman expansion during the early republic is poor, but the fact that Rome created 14 new rustic tribes during the years 387–241 BC suggests that population growth could have been a driving force. Furthermore, Romans living on the frontier may have strongly favoured war against restless neighbours, such as Gauls and Samnites. The animal husbandry of the latter involved seasonal migrations between summer uplands and winter lowlands, which caused friction between them and settled Roman farmers.

Motives for Roman imperialism

Though the Romans did not wage wars for religious ends, they often used religious means to assist their war effort. The fetial priests were used for the solemn official declaration of war. According to fetial law, Rome could enjoy divine favour only if it waged just wars—that is, wars of self-defense. In later practice, this often simply meant that Rome maneuvered other states into declaring war upon it. Then Rome followed with its declaration, acting technically in self-defense; this strategy had the effect of boosting Roman morale and sometimes swaying international public opinion.

Rome's first major war against an organized state was fought with Fidenae (437–426 BC), a town located just upstream from Rome. After it had been conquered, its land was annexed to Roman territory. Rome next fought a long and difficult war against Veii, an important Etruscan city not far from Fidenae. Later Roman historians portrayed the war as having lasted 10 years (406–396 BC), patterning it after the mythical Trojan War of the Greeks. After its conquest, Veii's tutelary goddess, Queen Juno, was solemnly summoned to Rome. The city's territory was annexed, increasing Roman territory by 84 percent and forming four new rustic tribes. During the wars against Fidenae and Veii, Rome increased the number of military tribunes with consular power from three to four and then from four to six. In 406 BC Rome instituted military pay, and in 403 BC it increased the size of its cavalry. The conquest of Veii opened southern Etruria to further Roman expansion. During the next few years, Rome proceeded to found colonies at Nepet and Sutrium and forced the towns of Falerii and Capena to become its allies. Yet, before Roman strength increased further, a marauding Gallic tribe swept down from the Po River valley, raided Etruria, and descended upon Rome. The Romans were defeated in the battle of the Allia River in 390 BC, and the Gauls captured and sacked the city; they departed only after they had received ransom in gold. Henceforth the Romans greatly feared and respected the potential strength of the Gauls. Later Roman historians, however, told patriotic tales about the commanders Marcus Manlius and Marcus Furius Camillus in order to mitigate the humiliation of the defeat.

Roman power had suffered a great reversal, and 40 years of hard fighting in Latium and Etruria were required to restore it fully. The terms of the second treaty between Rome and Carthage (348 BC) show Rome's sphere of in-



Roman expansion in Italy from 298 to 201 BC.

fluence to be about the same as it had been at the time of the first treaty in 509, but Rome's position in Latium was now far stronger.

The Samnite Wars. During the 40 years after the second treaty with Carthage, Rome rapidly rose to a position of hegemony in Italy south of the Po valley. Much of the fighting during this time consisted of three wars against the Samnites, who initially were not politically unified but coexisted as separate Oscan-speaking tribes of the central and southern Apennines. Rome's expansion was probably responsible for uniting these tribes militarily to oppose a common enemy. Both the rugged terrain and the tough Samnite soldiers proved to be formidable challenges, which forced Rome to adopt military innovations that were later important for conquering the Mediterranean.

The First Samnite War

Despite its brevity (343–341 BC), the First Samnite War resulted in the major acquisition to the Roman state of the rich land of Campania with its capital of Capua. Roman historians modeled their description of the war's beginning on the Greek historian Thucydides' account of the outbreak of the Peloponnesian War between Athens and Sparta. Nevertheless, they were probably correct in stating that the Campanians, when fighting over the town of Capua with the Samnites, allied themselves with Rome in order to utilize its might to settle the quarrel. If so, this may have been the first of many instances in which Rome went to war after being invited into an alliance by a weaker state already at war. Once invited in, Rome usually absorbed the allied state after defeating its adversary. In any event, Campania now somehow became firmly attached to Rome; it may have been granted Roman citizenship without the right to vote in Rome (*civitas sine suffragio*). Campania was a major addition to Rome's strength and manpower.

The absorption of Campania provoked the Latins to take up arms against Rome to maintain their independence. Since the Gallic sack of Rome in 390 BC, the city had become increasingly dominant within the Latin League. In 381 BC Tusculum was absorbed by being given Roman citizenship. In 358 BC Rome created two more rustic tribes from territory captured along the Volscian coast. The Latin War (340–338 BC) was quickly decided in Rome's favour. Virtually all of Latium was given Roman citizenship and became Roman territory, but the towns retained their local governments. The large states of Praeneste and Tibur maintained nominal independence by becoming Rome's military allies. Thus the Latin League was abolished; but the legal rights that the Latins had enjoyed among themselves were retained by Rome as a legal status, the Latin right (*ius Latii*), and used for centuries as an intermediate step between non-Roman status and full Roman citizenship.

Rome was now the master of central Italy and spent the next decade organizing and pushing forward its frontier through conquest and colonization. The Romans soon confronted the Samnites of the middle Liris (modern Liri) River valley, sparking the Second, or Great, Samnite War (326–304 BC). During the first half of the war Rome suffered serious defeats, but the second half saw Rome's recovery, reorganization, and ultimate victory. In 321 BC a Roman army was trapped in a narrow canyon near the Caudine Forks and compelled to surrender, and Rome was forced to sign a five-year treaty. Later Roman historians, however, tried to deny this humiliation by inventing stories of Rome's rejection of the peace and its revenge upon the Samnites. In 315 BC, after the resumption of hostilities, Rome suffered a crushing defeat at Lautulae. Ancient sources state that Rome initially borrowed hoplite tactics from the Etruscans (used during the 6th or 5th centuries BC) but later adopted the manipular system of the Samnites, probably as a result of Samnite success at this time. The manipular formation resembled a checkerboard pattern, in which solid squares of soldiers were separated by empty square spaces. It was far more flexible than the solidly massed hoplite formation, allowing the army to maneuver better on rugged terrain. The system was retained throughout the republic and into the empire. During these same years Rome organized a rudimentary navy, constructed its first military roads (construction of the Via Appia was begun in 312 BC and of the Via Valeria in 306), and increased the size of its annual military levy as seen from the increase of annually elected military tribunes from 6 to 16. During the period 334–295 BC, Rome founded 13 colonies against the Samnites and created six new rustic tribes in annexed territory. During the last years of the war, the Romans also extended their power into northern Etruria and Umbria. Several successful campaigns forced the cities in these areas to become Rome's allies. The Great Samnite War finally ended in Rome's victory. During the final phase of this war, Rome, on another front, concluded its third treaty with Carthage (306 BC), in which the Carthaginians acknowledged all of Italy as Rome's sphere of influence.

The manipular formation

The Third Samnite War (298–290 BC) was the last desperate attempt of the Samnites to remain independent. They persuaded the Etruscans, Umbrians, and Gauls to join them. Rome emerged victorious over this formidable coalition at the battle of Sentinum in 295 and spent the remainder of the war putting down lingering Samnite resistance. They henceforth were bound to Rome by a series of alliances.

The Pyrrhic War, 280–275 BC. Rome spent the 280s BC putting down unrest in northern Italy, but its attention was soon directed to the far south as well by a quarrel between the Greek city of Thurii and a Samnite tribe. Thurii called upon the assistance of Rome, whose naval operations in the area provoked a war with the Greek city of Tarentum. As in previous conflicts with Italian peoples, Tarentum summoned military aid from mainland Greece, calling upon King Pyrrhus of Epirus, one of the most brilliant generals of the ancient world. Pyrrhus arrived in southern Italy in 280 BC with 20 elephants and 25,000 highly trained soldiers. After defeating the Romans at Heraclea and stirring up revolt among the Samnites, he offered peace terms that would have confined Roman power to central Italy. When the Senate wavered, Appius Claudius, an aged blind senator, roused their courage and persuaded them to continue fighting. Pyrrhus again defeated the Romans in 279 at Asculum. His losses in the two battles numbered 7,500 (almost one-third of his entire force). When congratulated on his victory, Pyrrhus, according to Plutarch, "replied . . . that one other such would utterly undo him." This type of victory has since been referred to as Pyrrhic victory. Pyrrhus then left Italy and aided the Greeks of Sicily against Carthage; he eventually returned to Italy and was defeated by the Romans in 275 BC at Beneventum. He then returned to Greece, while Rome put down resistance in Italy and took Tarentum itself by siege in 272.

Rome was now the unquestioned master of Italy. Roman territory was a broad belt across central Italy, from sea to sea. Latin colonies were scattered throughout the peninsula. The other peoples of Italy were bound to Rome by a series of bilateral alliances that obligated them to provide Rome with military forces in wartime. According to the Roman census of 225 BC, Rome could call upon 700,000 infantry and 70,000 cavalry from its own citizens and allies. The conquest of Italy engendered a strong military ethos among the Roman nobility and citizenry, provided Rome with considerable manpower, and forced it to develop military, political, and legal institutions and practices for conquering and absorbing foreign peoples. The Pyrrhic War demonstrated that Rome's civilian army could wage a successful war of attrition against highly skilled mercenaries of the Mediterranean world.

(G.E.Fo.)

THE MIDDLE REPUBLIC (264–133 BC)

The first two Punic Wars. Rome's rapidly expanding sphere of hegemony brought it almost immediately into conflict with non-Italian powers. In the south, the main opponent was Carthage. In violation of the treaty of 306, which (historians tend to believe) had placed Sicily in the Carthaginian sphere of influence, Rome crossed the straits of Messana (between Italy and Sicily) embarking on war. (Rome's wars with Carthage are known as the "Punic Wars"; the Romans called the Carthaginians Poeni



A Roman war galley with infantry on deck. In the Vatican Museums and Galleries.
Alinari/Art Resource, New York City

[Phoenicians], from which derived the adjective "Punic.")

First Punic War (264–241 BC). The proximate cause of the first outbreak was a crisis in the city of Messana (Messina). A band of Campanian mercenaries, the Mamertini, who had forcibly established themselves within the town and were being hard pressed in 264 by Hieron II of Syracuse, applied for help to both Rome and Carthage. The Carthaginians, arriving first, occupied Messana and effected a reconciliation with Hieron. The Roman commander, nevertheless, persisted in forcing his troops into the city; he succeeded in seizing the Carthaginian admiral during a parley and induced him to withdraw. This aggression involved Rome in war with Carthage and Syracuse.

Operations began with their joint attack upon Messana, which the Romans easily repelled. In 263 the Romans advanced with a considerable force into Hieron's territory and induced him to seek peace and alliance with them. In 262 they besieged and captured the Carthaginian base at Agrigentum on the south coast of the island. The first years of the war left little doubt that Roman intentions extended beyond the protection of Messana.

In 260 the Romans built their first large fleet of standard battleships. At Mylae (Milazzo), off the north Sicilian coast, their admiral Gaius Duilius defeated a Carthaginian squadron of more maneuverable ships by grappling and boarding. This left Rome free to land a force on Corsica (259) and expel the Carthaginians, but it did not suffice to loosen their grasp on Sicily. A large Roman fleet sailed out in 256, repelled the entire Carthaginian fleet off Cape Ecnomus (near modern Licata), and established a fortified camp on African soil at Clypea (Kélibia in Tunisia). The Carthaginians, whose citizen levy was utterly disorganized, could neither keep the field against the invaders nor prevent their subjects from revolting. After one campaign they were ready to sue for peace, but the terms offered by the Roman commander Marcus Atilius Regulus were intolerably harsh. Accordingly, the Carthaginians equipped a new army in which cavalry and elephants formed the strongest arm. In 255 they offered battle to Regulus, who had taken up position with an inadequate force near Tunis, outmaneuvered him, and destroyed the bulk of his army. A second Roman fleet, which reached Africa after defeating the full Carthaginian fleet off Cape Hermaeum (Cape Bon), withdrew all the remaining troops.

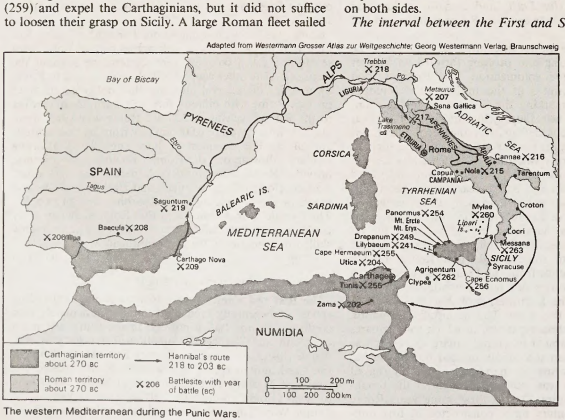
The campaign against Sicily

The Romans now directed their efforts once more against Sicily. In 254 they captured the important fortress of Panormus (Palermo), but when Carthage moved reinforcements onto the island, the war again came to a standstill. In 251 or 250 the Roman general Caecilus Metellus at last staged a pitched battle near Panormus, in which the enemy's force was effectively crippled. This victory was followed by a siege of the chief Punic base at Lilybaeum (Marsala), together with Drepanum (Trapani), by land and sea. In the face of resistance, the Romans were compelled to withdraw in 249; in a surprise attack upon Drepanum the Roman fleet under the command of admiral Publius Claudius Pulcher lost 93 ships. This was the Romans' only naval defeat in the war. Their fleet, however, had suffered a series of grievous losses by storm and was now so reduced that the attack upon Sicily had to be suspended. At the same time, the Carthaginians, who felt no less severely the financial strain of the prolonged struggle, reduced their forces and made no attempt to deliver a counterattack.

In 242 Rome resumed operations at sea. A fleet of 200 warships was equipped and sent out to renew the blockade of Lilybaeum. The Carthaginians hastily assembled a relief force, but in a battle fought off the Aegates, or Aegusae (Aegadian) Islands, west of Drepanum, their fleet was caught at a disadvantage and was largely sunk or captured (March 10, 241). This victory, by giving the Romans undisputed command of the sea, rendered certain the ultimate fall of the Punic strongholds in Sicily. The Carthaginians accordingly opened negotiations and consented to a peace by which they ceded Sicily and the Lipari Islands to Rome and paid an indemnity of 3,200 talents. The protracted nature of the war and the repeated loss of ships resulted in an enormous loss of life and resources on both sides.

The interval between the First and Second Punic Wars

Adapted from *Westermann Grosser Atlas zur Weltgeschichte*: Georg Westermann Verlag, Braunschweig



The western Mediterranean during the Punic Wars.

(241–218 BC). The loss of naval supremacy not only deprived the Carthaginians of their predominance in the western Mediterranean but exposed their overseas empire to disintegration under renewed attacks by Rome. Even the Greek historian Polybius, an admirer of Rome, considered the subsequent Roman actions against Carthage aggressive and unjustified. A gross breach of the treaty was perpetrated when a Roman force was sent to occupy Sardinia, whose insurgent garrison had offered to surrender the island (238). To the remonstrances of Carthage the Romans replied with a declaration of war and only withheld their attack upon the cession of Sardinia and Corsica and the payment of a further indemnity.

From this episode it became clear that Rome intended to use the victory to the utmost. To avoid further infringement of its hegemony, Carthage had little choice but to respond with force. The recent complications of foreign and internal strife had indeed so weakened the Punic power that the prospect of renewing the war under favourable circumstances seemed remote. Yet Hamilcar Barca sought to rebuild Carthaginian strength by acquiring a dominion in Spain where Carthage might gain new wealth and manpower. Invested with an unrestricted foreign command, he spent the rest of his life founding a Spanish empire (237–228). His work was continued by his son-in-law Hasdrubal and his son Hannibal, who was placed at the head of the army in 221. These conquests aroused the suspicions of Rome, which in a treaty with Hasdrubal confined the Carthaginians to the south of the Ebro River. At some point Rome also entered into relations with Saguntum (Sagunto), a town on the east coast, south of the Ebro. To the Carthaginians it seemed that once again Rome was expanding its interests into their sphere of hegemony. In 219 Hannibal laid siege to Saguntum and carried the town in spite of stubborn defense. The Romans responded with an ultimatum demanding that the Carthaginians surrender Hannibal or go to war. The Carthaginian council supported Hannibal and accepted the war.

Second Punic War (218–201 BC). It seemed that the superiority of the Romans at sea ought to have enabled them to choose the field of battle. They decided to send one army to Spain and another to Sicily and Africa. But before their preparations were complete, Hannibal began the series of operations that dictated the course of the war for the greater part of its duration. He realized that as long as the Romans commanded the resources of an undivided Italian confederacy, no foreign attack could overwhelm them beyond recovery. Thus he conceived the plan of cutting off their source of strength by carrying the war into Italy and causing a disruption of the league. His chances of ever reaching Italy seemed small, for the sea was guarded by the Roman fleets and the land route was long and arduous.

Hannibal's strategy

But the very boldness of his enterprise contributed to its success; after a six months' march through Spain and Gaul and over the Alps, which the Romans were nowhere in time to oppose, Hannibal arrived (autumn 218) in the plain of the Po with 20,000 foot soldiers and 6,000 horses, the pick of his African and Spanish levies. At the end of the year, Hannibal, by superior tactics, repelled a Roman army on the banks of the Trebbia River, inflicting heavy losses, and thus made his position in northern Italy secure.

In 217 the campaign opened in Etruria, into which the invading army, largely reinforced by Gauls, penetrated via an unguarded pass. A rash pursuit by the Roman field force led to its being entrapped on the shore of Lake Trasimene (Trasimeno) and destroyed with a loss of at least 15,000 men. This catastrophe left Rome completely uncovered; but Hannibal, having resolved not to attack the capital before he could collect a more overwhelming force, directed his march toward the south of Italy, where he hoped to stir up the peoples who had formerly been the most stubborn enemies of Rome. The Italians, however, were slow everywhere to join the Carthaginians. A new Roman army under the dictator Quintus Fabius Maximus ("Cunctator") dogged Hannibal's steps on his forays through Apulia and Campania and prevented him from acquiring a permanent base of operations.

The eventful campaign of 216 was begun by a new, aggressive move on the part of Rome. An exceptionally strong field army, variously estimated at between 48,000 and 85,000 men, was sent to crush the Carthaginians in open battle. On a level plain near Cannae in Apulia, Hannibal deliberately allowed his centre to be driven in by the numerically superior Romans, while Hasdrubal's cavalry wheeled around so as to take the enemy in flank and rear. The Romans, surrounded on all sides, were practically annihilated, and the loss of citizens was perhaps greater than in any other defeat that befell the republic.

The Battle of Cannae

The effect of the battle on morale was no less momentous. The southern Italian peoples seceded from Rome, the leaders of the movement being the people of Capua, at the time the second greatest town of Italy. Reinforcements were sent from Carthage, and several neutral powers prepared to throw their weight into the scale on Hannibal's behalf. But the great resources of Rome, though terribly reduced in respect to both men and money, were not yet exhausted. In northern and central Italy the insurrection spread but little and could be sufficiently guarded against with small detachments. In the south the Greek towns of the coast remained loyal, and the numerous Latin colonies continued to render important service by interrupting free communication between the rebels and detaining part of their forces.

In Rome itself the crisis gave way to a unanimity unparalleled in the annals of the republic. The guidance of operations was henceforth left to the Senate, which, by maintaining a persistent policy until the conflict was brought to a successful end, earned its greatest title to fame. But it also produced a severe strain, released through cruel religious rites, which were an embarrassment to later Roman authors. The disasters were interpreted as evidence of divine wrath at Roman impiety, to be propitiated by punishment (burial alive) of two offending Vestal Virgins and by the human sacrifice of a Gallic and Greek man and woman.

The subsequent campaigns of the war in Italy assumed a new character. Though the Romans contrived at times to raise 200,000 men, they could spare only a moderate force for field operations. Their generals, among whom the veterans Fabius and Marcus Claudius Marcellus frequently held the most important commands, rarely ventured to engage Hannibal in the open and contented themselves with observing him or skirmishing against his detachments. Hannibal, whose recent accessions of strength were largely discounted by the necessity of assigning troops to protect his new allies or secure their wavering loyalty, was still too weak to undertake a vigorous offensive. In the ensuing years the war resolved itself into a multiplicity of minor engagements, which need not be followed in detail. In 216 and 215 the chief seat of war was Campania, where Hannibal, vainly attempting to establish himself on the coast, experienced a severe repulse at Nola.

In 214 the main Carthaginian force was transferred from Apulia in hopes of capturing Tarentum (Taranto), a suitable harbour by which Hannibal might have secured his overseas communications. In 213–212 the greater part of Tarentum and other cities of the southern seaboard at last came into Hannibal's power, but in the meantime the Romans were suppressing the revolt in Campania and in 212 were strong enough to place Capua under blockade. They severely defeated a Carthaginian relief force and could not be permanently dislodged even by Hannibal himself. In 211 Hannibal made a last effort to relieve his allies by a feint upon Rome itself, but the besiegers refused to be drawn away from their entrenchments, and eventually Capua was starved into surrender. The Romans in 209 gained a further important success by recovering Tarentum. Though Hannibal still won isolated engagements, he was slowly being driven back into the extreme south of the peninsula.

The campaign of 207

In 207 the arrival of a fresh invading force produced a new crisis. Hasdrubal, who in 208–207 had marched overland from Spain, appeared in northern Italy with a force scarcely inferior to the army that his brother had brought in 218. After levying contingents of Gauls and Ligurians, he marched down the east coast with the object of joining his brother in central Italy for a direct attack upon Rome

itself. By this time the steady drain of men and money was telling so severely upon the confederacy that some of the most loyal allies protested their inability to render further help. Nonetheless, by exerting a supreme effort, the Romans raised their war establishment to the highest total yet attained and sent a strong field army against each Carthaginian leader. Before reaching Hannibal, Hasdrubal was met in northern Italy by the army of Marcus Livius Salinator, reinforced by part of Gaius Claudius Nero's army. The battle on the banks of the Metaurus (Metauro) River was evenly contested until Nero, with a dexterous flanking movement, cut off the enemy's retreat. The bulk of Hasdrubal's army was destroyed, and he himself was killed. His head was tossed into his brother's camp as an announcement of his defeat.

The campaign of 207 decided the war in Italy. Though Hannibal still maintained himself for some years in southern Italy, this was chiefly due to the exhaustion of Rome. In 203 Hannibal, in accordance with orders received from home, sailed back to Africa; and another expedition under his brother Mago, which had sailed to Liguria in 205 and endeavoured to rouse the slumbering discontent of the people in Cisalpine Gaul and Etruria, was forced to withdraw.

Campaigns in Sicily and Spain. Concurrently with the great struggle in Italy, the Second Punic War was fought on several other fields. To the east King Philip V of Macedon began the First Macedonian War (214–205) in concert with the Carthaginians, when the Roman power seemed to be breaking up after Cannae. Although this compelled the Romans to stretch their already severely strained resources still further by sending troops to Greece, the diversions Roman diplomacy provided for Philip in Greece and the maintenance of a Roman patrol squadron in the Adriatic Sea prevented any effective cooperation between Philip and Hannibal.

Agriculture in Italy had collapsed, and the Romans had to look to Sardinia and Sicily for their food supply. Sardinia was attacked by Carthaginians in 215, but a small Roman force was enough to repel the invasion. In Sicily the death of Hieron II, Rome's steadfast friend, in 215 left the realm of Syracuse to his inexperienced grandson Hieronymus. The young prince abruptly broke with the Romans, but before hostilities commenced he was assassinated. The Syracusan people now repudiated the monarchy and resumed their republican constitution. When the Romans threatened terrible punishment, the Syracusans found it necessary to cooperate with the Carthaginians.

Roman attack on Syracuse

The Roman army and fleet under Marcus Claudius Marcellus, which speedily appeared before the town, were completely baffled by the mechanical contrivances that the Syracusan mathematician Archimedes had invented in 213 for the defense of the city. Meanwhile, the revolt against Rome spread in the interior of the island, and a Carthaginian fleet gained control of towns on the south coast. In 212 Marcellus at last broke through the defense of Syracuse and, in spite of the arrival of a Carthaginian relief force, took control of the whole town in 211. By the end of 210 Sicily was wholly under the power of Rome.

The conflict in Spain was second in importance only to the Italian war. From this country the Carthaginians drew large supplies of troops and money that might serve to reinforce Hannibal; hence it was in the interest of the Romans to challenge their enemy within Spain. Though the force that Rome at first spared for this war was small in numbers and rested entirely upon its own resources, the generals Publius Cornelius and Gnaeus Cornelius Scipio, by skillful strategy and diplomacy, not only won over the peoples north of the Ebro and defeated the Carthaginian leader Hasdrubal Barca in his attempts to restore communication with Italy but also carried their arms along the east coast into the heart of the enemy's domain.

But eventually the Roman successes were nullified by a rash advance. Deserted by their native contingents and cut off by Carthaginian cavalry, among which the Numidian prince Masinissa rendered conspicuous service, the Roman generals were killed and their troops destroyed (211).

Disturbances in Africa prevented the Punic commanders from exploiting their success. Before long the fall of Capua enabled Rome to transfer troops from Italy to Spain; and in 210 the best Roman general of the day, the young son and namesake of Publius Scipio, was placed in command by popular vote, despite his youth and lack of the prerequisite senior magistracies. He signalized his arrival by a bold and successful coup de main upon the great arsenal of Carthago Nova (Cartagena) in 209. Though after an engagement at Baecula (Bailen; 208) he was unable to prevent Hasdrubal Barca from marching away to Italy, Scipio profited by his opponent's departure to push back the remaining hostile forces the more rapidly. A last effort by the Carthaginians to retrieve their losses with a fresh army was frustrated by a great Roman victory at Ilipa, near Seville, and by the end of the year 206 the Carthaginians had been driven out of Spain.

The Battle of Ilipa

The war in Africa. In 205 Scipio, who had returned to Rome to hold the consulship, proposed to follow up his victories by an attack on the home territory of Carthage. Though the presence of Hannibal in Italy deterred Fabius and other senators from sanctioning this policy, Scipio gradually overbore all resistance. He built up a force, which he organized and supplemented in Sicily, and in 204 sailed across to Africa. He was met there by a combined levy of Carthage and King Syphax of Numidia and for a time was penned to the shore near Utica. But in the spring he extricated himself by a surprise attack on the enemy's camp, which resulted in the total loss of the allied force by sword or fire.

In the campaign of 203 a new Carthaginian force was destroyed by Scipio on the Great Plains 75 miles from Utica, their ally Syphax was captured, and the renegade Masinissa was reinstated in the kingdom from which Syphax had recently expelled him. These disasters induced the Carthaginians to sue for peace; but before the very moderate terms that Scipio offered could be definitely accepted, a sudden reversal of opinion caused them to recall Hannibal's army for a final trial of war and to break off negotiations. In 202 Hannibal assumed command of a composite force of citizen and mercenary levies reinforced by a corps of his veteran Italian troops.

After negotiations failed, Scipio and Hannibal met in the Battle of Zama. Scipio's force was somewhat smaller in numbers but well trained throughout and greatly superior in cavalry. His infantry, after evading an attack by the Carthaginian elephants, cut through the first two lines of the enemy but was unable to break the reserve corps of Hannibal's veterans. The battle was ultimately decided by the cavalry of the Romans and their new ally Masinissa, who by a maneuver recalling the tactics of Cannae took Hannibal's line in the rear and destroyed it.

The Carthaginians again applied for peace and accepted the terms that Scipio offered. They were compelled to cede Spain and the Mediterranean islands still in their hands, to surrender their warships, to pay an indemnity of 10,000 talents within 50 years, and to forfeit their independence in affairs of war and foreign policy.

The Second Punic War, by far the greatest struggle in which either power engaged, had thus ended in the complete triumph of Rome, although not because of any faultiness in the Carthaginians' method of attack. Carthage could only hope to win by invading Italy and using the enemy's home resources against him. The failure of Hannibal's brilliant endeavour was ultimately due to the stern determination of the Romans and to the nearly inexhaustible manpower from their Italian confederacy, which no shock of defeat or strain of war could entirely disintegrate. Although Rome and its allies suffered casualties of perhaps one-fifth of their adult male population, they continued fighting. For Polybius, the Second Punic War illustrated the superiority of the strong Roman constitution over Hannibal's individual genius.

(M.Car./H.H.S./R.P.Sa.)

The establishment of Roman hegemony in the Mediterranean world. *Roman expansion in the eastern Mediterranean.* Just before the Second Punic War, Rome had projected its power across the Adriatic Sea against the Illyrians. As noted, Philip V of Macedon in turn had joined the Carthaginians for a time during the war in an attempt to stem the tide of Roman expansion but had

agreed to terms of peace with Rome's allies, the Aetolians, in 206 and then with Rome in the Peace of Phoenice of 205. Immediately after the Second Punic War the Roman Senate moved to settle affairs with Philip, despite the war-weary centuriate assembly's initial refusal to declare war. Historians have debated Rome's reasons for this momentous decision, with suggestions ranging from a desire to protect Athenians and other Greeks from Philip out of philhellenism to fear of a secret alliance between Philip and the Seleucid king Antiochus III. Yet these suggestions are belied by the fact that Rome later treated the Greek cities callously and that no fear is apparent in Rome's increasing demands on Philip and in its refusal to negotiate seriously with him through the course of the war. Rather, the Second Macedonian War (200–196) fits the long pattern of Roman readiness to go to war in order to force ever more distant neighbours to submit to superior Roman power.

The Second Macedonian War

In the winter of 200–199, Roman legions marched into the Balkans under the command of Publius Sulpicius Galba. During the next two years there was no decisive battle, as the Romans gathered allies among the Greeks—not only their previous allies, the Aetolians, but also Philip's traditional allies, the Achaeans, who recognized Roman military superiority. The consul of 198, Titus Quinctius Flamininus, took over the command and defeated Philip at the battle of Cynoscephalae in 197. The terms of settlement allowed Philip to remain king of Macedon but stipulated payment of an indemnity and restrictions on campaigning beyond the borders of his kingdom. Flamininus then sought to win the goodwill of the Greeks with his famous proclamation of their liberation at the Isthmian Games of 196. To lend credibility to this proclamation, he successfully argued against senatorial opposition for the withdrawal of Roman troops from all Greece, including the strategically important "Fetters" (the key garrisons of Acrocorinth, Chalcis, and Demetrias).

Even before the Romans withdrew, the seeds had been sown for their reentry into the East. As an active king, Antiochus III set out to recover the ancestral possessions of his kingdom on the western coast of Anatolia and in Thrace. In response to the Roman demand that he stay out of Europe, the king attempted to negotiate. When the Romans showed little interest in compromise, Antiochus accepted the invitation of Rome's former allies, the Aetolians, who felt they had not been duly rewarded with additional territory after the victory over Philip, to liberate the Greeks. Upon crossing into Greece, however, the king found no enthusiasm among the other Greeks for a war of liberation and was defeated at Thermopylae in 191 by legions under the command of Manius Acilius Glabrio.

Antiochus returned home to gather a larger army. In 190 Lucius Cornelius Scipio was elected consul in Rome and was authorized to recruit a force for a campaign against Antiochus. Accompanying Lucius as a legate was his brother, the great general Scipio Africanus. In an attempt to avert war, Antiochus offered to accept the earlier Roman terms, only to find that the Romans had now extended their demands to keep Antiochus east of the Taurus Mountains of Anatolia. Unable to accept, Antiochus fought and lost to Scipio's army at Magnesia ad Sipylum in the winter of 190–189. In the following Treaty of Apamea (188), the Seleucid kingdom was limited to Asia east of the Taurus range and was required to pay an indemnity of 15,000 talents and to give up its elephants and all but 10 ships. Rome punished its opponents, the Aetolians, and rewarded its supporters, notably Pergamum and Rhodes, which were granted new territories, including Greek cities, at the expense of "the liberation of the Greeks." The consul of 189, Gnaeus Manlius Vulso, came east with reinforcements, took command of the legions, and proceeded to plunder the Galatians of Anatolia on the pretext of restoring order.

The withdrawal of Roman legions this time did not entail the withdrawal of a Roman presence from the Hellenistic East. On the contrary, according to Polybius, the Romans now "were displeased if all matters were not referred to them and if everything was not done in accordance with their decision." Continuing jealousies and disputes in the Greek world offered Rome opportunities to adjudicate and ultimately to intervene once again. In the Peloponnese the Achaean League was at odds with Sparta, wishing to bring Sparta into the league and to suppress the radical social program of its king, Nabis. Flamininus in 195 supported the independence of Sparta, but in 192 the Achaean leader, Philopoemen, induced Sparta to join the league with a promise of no interference in its internal affairs. When an infringement of the promise prompted the Spartans to secede, Philopoemen in 188 led an Achaean army to take Sparta, kill the anti-Achaean leaders, and force the city back into the league. Although the Senate heard complaints, it took no immediate action. Then, in 184, the Senate reasserted its own terms for settlement but was circumvented by Philopoemen, who reached a separate agreement with the Spartans. The independent-minded Philopoemen died the following year in a campaign by the league to suppress a revolt of Messene. His death led to a change of leadership, as the pro-Roman Callicrates (regarded by Polybius as a sycophant) began a policy of obeying Rome's every wish.

Philopoemen and the Achaean League

Meanwhile, tensions between Rome and Philip were increasing. Philip had supported Rome's war with Antiochus in the hope of recovering Thessalian and Thracian territory, but in this he was disappointed by the Romans. They did, however, return Philip's younger son, Demetrius, taken to Rome as a hostage in 197—a reward with tragic consequences. During his years as a hostage, Demetrius had made senatorial friendships, which aroused suspicions at home that the Romans would prefer to see Demetrius rather than his elder brother, Perseus, succeed Philip. Philip ordered the death of Demetrius in 181 and then died in 179, leaving his throne to Perseus, the last king of Macedon.

Perseus' activism started a stream of complaints to the Senate from neighbouring Greek powers from 175 onward. The king's real intentions are unclear; perhaps Polybius was right that he wished to make the Romans "more cautious about delivering harsh and unjust orders to Macedonians." The Senate listened to the unfavourable interpretations of Perseus' enemies, who claimed that the king's actions revealed an intent to attack Rome. Like his father, Perseus campaigned to extend Macedonian power to the northeast and south and marched through Greece as far as Delphi. He solicited alliances with the Achaean League and other Greek states, which some of the leaders hostile to Rome would have liked to accept. He arranged dynastic marriages with other Hellenistic kings, taking the daughter of Seleucus IV as his wife and giving the hand of his sister to Prusias II of Bithynia. Although these actions could have been viewed as the behaviour expected of a Hellenistic monarch, Eumenes of Pergamum suggested to the Senate that Perseus was preparing for war against Rome. After the Senate decided on war, it sent Quintus Marcius Philippus to propose a truce and to give Perseus false hopes of negotiation in order to allow the consul of 171, Publius Licinius Crassus, to land his army on the Illyrian coast unhindered—a ploy decried by some older senators as "the new wisdom."

Perseus' initial success against the Roman army in Thessaly in 171 did not alter the massive imbalance of power; the Romans again refused the king's offer to negotiate. Over the next three years Roman commanders devoted more effort to plunder than to the defeat of Perseus. In a notorious incident, the praetor Lucius Hortensius anchored his fleet at Abdera, a city allied with Rome, and demanded supplies; when the Abderitans asked to consult the Senate, Hortensius sacked the town, executed the leading citizens, and enslaved the rest. When complaints reached the Senate, weak attempts were made to force the Roman commanders to make restitution. In 168 the experienced Lucius Aemilius Paullus was reelected consul and sent out to restore discipline. He quickly brought the Third Macedonian War to an end by defeating Perseus in the Battle of Pydna in June 168. Perseus was deposed, and Macedonia was divided into four republics, which were forbidden to have relations with one another; they paid tribute to Rome at half the rate they had previously paid to the king.

The defeat of Perseus

In 167 Rome proceeded to punish those who had sided with Perseus (such as the Illyrian Genthius), those whose loyalty had wavered (such as Eumenes), and even those who had contemplated acting as mediators in the war (such as the Rhodians). In Illyria, Paullus, on instructions from the Senate, swept through the countryside enslaving 150,000 inhabitants from 70 Epirote towns. In Achaea, 1,000 leading men suspected of Macedonian sympathies were taken as hostages to Rome. (Among them was Polybius, who befriended the noble Scipionic family and wrote his great history of the rise of Rome with the aid of privileged access to the views of the senatorial leadership.) Eumenes was refused a hearing before the Senate on his visit to Italy; his fall from favour prompted his enemies to dispute his territory, and in 164 a Roman embassy in Anatolia publicly invited complaints against the king. Rhodes had thrived as the leading trade centre of the eastern Mediterranean, using its considerable resources to control piracy; now Rome undermined its economy and power by making the island of Delos a free port, thereby depriving Rhodes of its income from harbour dues. Territory in Lycia and Caria on the mainland, granted to Rhodes in 189, was now taken away. But the far harsher proposal in the Senate to declare Rhodes an enemy and to destroy it was opposed by senior senators such as Cato the Censor and was voted down. As a result of the weakening of Rhodes, piracy became rampant in the eastern Mediterranean (the young Julius Caesar was captured by pirates). During the next century Roman senators did not find the political will to suppress the piracy, perhaps in part because it served their interests; pirates supplied tens of thousands of slaves for their Italian estates and disrupted the grain trade, thus raising prices for their produce in Rome.

The rise of piracy

The arrangements of 167 served the Roman policy of weakening the powers of the eastern Mediterranean. In the previous year Rome had also intervened to stop Seleucid expansion into Egypt. In a famous episode, the Roman ambassador Gaius Popillius Laenas delivered to Antiochus IV the Senate's demand that the king withdraw from Egypt. When the king requested time for consultation, Popillius "drew a circle around the king with a stick he was carrying and told him not to leave the circle until he gave his response. The king was astonished at this occurrence and the display of superiority, but, after a brief time, said he would do all the Romans demanded."

The power vacuum fostered by the Romans was not ultimately conducive to stability. An adventurer, Andriscus, claiming to be descended from the Macedonian dynasty, was able to enter the Macedonian republics without serious resistance. He was successful enough in raising an army to defeat the first Roman force sent against him in 149 under the command of the praetor Publius Iuventius Thalna (who was killed). A second Roman army under Quintus Caecilius Metellus defeated the pretender in 148. With the death of Callicrates, leadership of the Achaean League passed to Critolaus and Diaeus, outspoken proponents of Greek independence from Rome. In 147 a Roman embassy was sent to intervene in the affairs of the league by supporting the secession of Sparta and also by calling for the detachment of Corinth and Argos from the league. The embassy provoked a violent reply. When further negotiations were blocked by Critolaus, Rome declared war on the Achaeans in 146, citing as reason the ill-treatment of their embassy. Metellus (now with the appellation of "Macedonicus"), having delayed with his army, marched against Critolaus and defeated him in Locris. Then Lucius Mummius Archaicus, consul of 146, took over the command and defeated Diaeus and the remaining Achaeans. The Senate ordered Mummius to teach a lesson to the Greeks: the venerable city of Corinth was sacked, its treasures taken to Rome, and its buildings burned to the ground.

The nature of Roman domination in the East began to change decisively after these wars: in place of influence through embassies, arbitration of disputes, and the occasional military incursion came direct rule. Macedonia was annexed as a province, to be governed and taxed by a Roman proconsul, who also watched over the Greek cities to the south, where the leagues were disbanded.

Direct Roman rule in the East

Farther east, the kingdom of Pergamum was added as the province of Asia, as a bequest to the Roman people from Attalus III in 133.

Roman expansion in the western Mediterranean. If Roman military intervention in the east was sporadic in the 2nd century, campaigning in northern Italy and Spain was nearly continuous. During Hannibal's invasion of Italy, the Insubres and Boii, Gallic peoples in the Po valley, had joined the Carthaginians against Rome. In 200 the Gauls and Ligurians combined forces and sacked the Latin colony of Placentia in an attempt to drive the Romans out of their lands. In the following years consular armies repeatedly attacked the Gauls. In 194 Lucius Valerius Flaccus won a decisive victory over the Insubres; in 192 the leading Boii under severe pressure went over to the Roman side, signaling the coming defeat of their tribe. Following their victories, the Romans sent thousands of new colonists to the Po valley to reinforce the older colonies of Placentia and Cremona (190) and to establish new colonies, notably Bononia (189) and Aquileia (181).

During the same period the Romans were at war with the Ligurian tribes of the northern Appennines. The serious effort began in 182, when both consular armies and a proconsular army were sent against the Ligurians. The wars continued into the 150s, when victorious generals celebrated two triumphs over the Ligurians. Here also the Romans drove many natives off their land and settled colonies in their stead (*e.g.,* Luna and Luca in the 170s).

As a result of the Second Punic War, Roman legions had marched into Spain against the Carthaginians and remained there after 201. The Romans formalized their rule in 197 by creating two provinces, Nearer and Further Spain. They also exploited the Spanish riches, especially the mines, as the Carthaginians had done. In 197 the legions were withdrawn, but a Spanish revolt against the Roman presence led to the death of one governor and required that the two praetorian governors of 196 be accompanied by a legion each. The situation was serious enough for the consul of 195, Cato the Censor, to be sent to Spain with two legions. From Cato comes the earliest extant firsthand account of Roman conquest. His comments show that he prided himself on his bravery and lack of greed as compared with other Roman commanders. Yet his narrative must overstate the extent and decisiveness of his success because fighting persisted for years to come, as later Roman governors sought to extend Roman control over more Spanish peoples—the Celtiberi of northeastern Spain, the Lusitani of modern-day Portugal, and the Vettones and Vaccaei of northwestern Spain. In 177 Tiiberius Sempronius Gracchus celebrated a triumph over the Celtiberi. The size of the Roman forces was probably then reduced from four to two legions; from 173 to 155 there was a lull in the regular campaigning. During these decades Spanish peoples brought complaints to Rome about corrupt governors.

Annual warfare resumed in Spain in 154, being perhaps in part a violent reaction to corrupt administration, and dragged on until 133. Labeled a "fiery war" (really wars), these struggles acquired a reputation for extreme cruelty; they brought destruction to the native population (*e.g.,* 20,000 Vaccaei were killed in 151 after giving themselves up to Lucius Licinius Lucullus) and made recruiting legionaries in Italy difficult. In Further Spain the Lusitanian leader Viriathus enjoyed some successes, including the surrender of a Roman army in 141–140 and a favourable treaty with Rome, but the next governor of the province, Quintus Servilius Caepio, arranged for his assassination in 139. Two years later in Nearer Spain, the Numantines also forced the surrender of an army under Gaius Hostilius Mancinus; the Senate later disavowed the agreement of equal terms and handed Mancinus, bound and naked, over to the Spaniards to absolve themselves of responsibility before the gods. The wars in Spain were brought to a conclusion in 133 by Publius Cornelius Scipio Aemilianus, who took Numantia after a long siege, enslaved the population, and razed the city.

A "fiery war"

It was Scipio Aemilianus (b. 185/184) who in the previous decade had imposed a similar final solution on Carthage in the Third Punic War (149–146). After the

Second Punic War, Carthage had recovered to the point that in 191 it offered to repay the remainder of the 50-year tribute of 200 talents per year in one lump sum. Rome's refusal of the offer suggests that beyond its monetary value the tribute had the symbolic importance of signifying subjection. Carthage's neighbour, the Numidian king Masinissa, had been granted as a reward for his support of Rome at the Battle of Zama his paternal kingdom and the western Numidian kingdom ruled by Syphax. During the next half century Masinissa periodically tried to exploit his favour in Rome by encroaching on Carthaginian territory. Initially, the Carthaginians submissively sought the arbitration of Rome in these disputes, but more often than not Roman judgment went in favour of Masinissa. After a series of losses, the Carthaginians in 151 decided to act on their own and raised an army to ward off the Numidian attacks. When a Roman delegation observed the Carthaginian army raised in breach of the treaty of 201, Rome was provided with the casus belli for a declaration of war in 149; Polybius, however, claims that the Senate had decided on this war "long before." The elderly Cato had been ending his speeches in the Senate since 153 with the notorious exhortation that "Carthage must be destroyed." Carthage desperately and pathetically tried to make amends, executing the generals of the expedition against the Numidians, surrendering to Rome, and handing over hostages, armour, and artillery. Only then did the Romans deliver their final demand: Carthage must be abandoned and the population moved to a new site inland. Such extreme terms could not be accepted.

The war against Carthage, with its prospects of rich booty, presented no recruiting problems for the Romans: huge land and naval forces were sent out under both consuls of 149, Lucius Marcius Censorinus and Manius Manilius. The imbalance of resources meant that the outcome was never in doubt, but the fortifications of Carthage delayed the Roman victory. The young Scipio Aemilianus was elected consul for 147, and by popular vote he was assigned the task of bringing the war to an end. He blockaded the city by land and sea, inflicting terrible suffering. Finally, in 146, the Roman army took Carthage, enslaved its remaining 50,000 inhabitants, burned the buildings to the ground, and ritually sowed the site with salt to guarantee that nothing would ever grow there again. Carthaginian territory was annexed as the province of Africa.

The destruction of Carthage

Explanations of Roman expansion. As one of the decisive developments in western history, Roman expansion has invited continual reinterpretation by historians. Polybius, who wrote his history in order to explain to other Greeks the reasons for Roman success, believed that after their victory over Hannibal the Romans conceived the aim of dominating all before them and set out to achieve it in the Second Macedonian War. If one accepts the Roman view that they fought only "just wars"—that is, only when provoked—then Roman conquest emerges as "one of the most important accidents in European history," as Rome had to defend itself from threats on all sides. Historians have suggested other motives for empire, such as a desire to profit from war, an interest in commercial expansion, or a love of the Greeks, who asked for protection against Hellenistic monarchs.

Major historical phenomena of this kind rarely receive final, decisive interpretations, but several assertions may be ventured. Some of the interpretations are anachronistic impositions on the ancient world; ancient testimony, for example, gives no support to commercial or mercantile explanations. Cultural and economic interpretations seem more appropriate. Roman culture placed a high value on success in war: *virtus* (courage and qualities of leadership) was displayed, above all, in war, and the triumph, a parade through Rome celebrating a major victory over an enemy, was the honour most highly prized by the senatorial generals who guided Roman decisions about war and peace. Moreover, these leaders, and the whole Roman people, were fully aware of the increasing profits of victory; in the 2nd century commanders and soldiers, as well as the city itself, were enriched by the glittering booty from Africa and the Greek East.

Yet, it is rightly pointed out, Roman intervention in the East was sporadic, not systematic, and the Romans did not annex territory in the Balkans, Anatolia, or North Africa for more than 50 years after their initial victories. The latter point, however, is not telling, since the Romans regarded defeated states allied to them as part of their *imperium*, whether or not they were under Roman provincial administration. The sporadic timing of the wars would seem to support the Romans' claim that they only reacted, justly, to provocations. But attention to the individual provocations should not blind the historian to the larger pattern of Roman behaviour. From 218 the Romans annually fielded major armies decade after decade. Rome was able to go to war every year in response to provocations only because it chose to define its interests and make alliances farther and farther afield. Polybius, as noted, reveals how the Romans were the masters of manipulation of circumstances to force opponents to behave in a way they could interpret as provocative. Therefore, the Roman interpretation of "just wars" and the Polybian interpretation of a universal aim to conquer need not be contradictory. The concept of "just war" may have justified any given war but does not explain the perpetual Roman readiness to go to war. For that the historian must look to Polybius' universal aim or to general political, social, economic, and cultural features of Rome. Finally, it must be remembered that in some instances it was clearly the Roman commander who provoked the war in order to plunder and to win a triumph (*e.g.*, Licinius Lucullus, governor of Nearer Spain, in 151).

The issue of "just wars"

Beginnings of provincial administration. Rome dominated its Latin and Italian neighbours by incorporating some into the Roman citizen body and by forming bilateral alliances with most of the Italian city-states. After the Punic Wars, Rome undertook to rule newly acquired territories directly as subject provinces. In 241 Sicily became Rome's first province, followed by Sardinia-Corsica in 238, and Spain, divided into two provinces, in 197. After a 50-year hiatus, Macedonia and Africa were annexed in 146, and the province of Asia (northwestern Anatolia) in 133. In principle, each province was to be administered in accordance with its *lex provinciae*, a set of rules drawn up by the conquering commander and a senatorial embassy. The *lex provinciae* laid down the organization of taxation, which varied from province to province.

The provincial administrative apparatuses were minimal and unprofessional, as the Romans relied heavily on the local elites as mediators. Each year a senatorial magistrate was sent out to govern with nearly unfettered powers. Because initially the governors were usually praetors, the addition of new provinces required the election of more praetors (increased to four in 227 and to six in 197). The assignments to provinces were done by lot. The governor took with him one of the quaestors to oversee the finances of provincial government and senatorial friends and relatives to serve as deputies and advisors (*legati*). Among the humbler functionaries assisting the governor were scribes to keep records and lictors with fasces (bundles of rods and axes) to symbolize gubernatorial authority and to execute sentences pronounced by the governor in criminal cases.

The governor's main duties were to maintain order and security and to collect revenues. The former often entailed command of an army to ward off external threats and to suppress internal disorders such as banditry. When not commanding his army, the governor spent his time hearing legal cases and arbitrating disputes. During the republic, revenue collection was left to private companies of *publicani*, so called because they won by highest bid the contract to collect the revenues. It was the governor's responsibility to keep the *publicani* within the bounds of the *lex provinciae* so that they did not exploit the helpless provincials too mercilessly, but this was difficult. Governors expected to make a profit from their term of office, and some collaborated with the *publicani* to strip the provinces of their wealth.

Gubernatorial duties

THE TRANSFORMATION OF ROME AND ITALY DURING THE MIDDLE REPUBLIC

The Greek historian Polybius admired Rome's balanced constitution, discipline, and strict religious observance as

the bases of the republic's success and stability. Yet Rome's very successes in the 2nd century undermined these features, leading to profound changes in the republic's politics, culture, economy, and society.

Citizenship and politics in the middle republic. The Romans organized their citizenry in a way that permitted expansion. This was regarded as a source of strength by contemporaries such as Philip V, who noted that Rome replenished its citizen ranks with freed slaves. The extension of citizenship continued in the early 2nd century, as in the grant of full citizen rights to Arpinum, Formiae, and Fundi in 188. Yet Rome's glittering successes made such openness ever more problematic. For one, the city attracted increasing numbers of Latins and allies, who wished to use their ancient right to migrate and take up Roman citizenship. The depletion of Latin and Italian towns prompted protests, until in 177 Rome took away the right of migration and forced Latin and Italian migrants to return to their hometowns to register for military service. Such measures were sporadically repeated in the following years. In addition, the flood of slaves into Rome from the great conquests increased the flow of foreign-born freedmen into the citizen body. Sempronius Gracchus (father of the famous tribunes) won senatorial approbation as censor in 168 by registering the freedmen in a single urban tribe and thus limiting their electoral influence. Despite these efforts, the nature and meaning of Roman citizenship were bound to change, as the citizen body became ever more diffuse and lived dispersed from Rome, the only place where the right of suffrage could be exercised.

Polybius greatly admired Rome's balanced constitution, with its elements of monarchy (magistrates), aristocracy (Senate), and democracy (popular assemblies). According to Greek political theory, each form of constitution was believed to be unstable and susceptible to decline until replaced by another. Yet Rome's system of balance, Polybius thought, was a check on the cycle of decline. By forcing the Roman constitution into the mold of Greek political theory, however, he exaggerated the symmetry of checks and balances. In reality, the Senate enjoyed a period of steady domination through the first two-thirds of the 2nd century, having emerged from the Second Punic War with high prestige. Only occasionally did the developing tensions and contradictions surface during these decades.

Competition for high office

Politics during the period was largely a matter of senatorial families competing for high office and the ensuing lucrative commands. Because offices were won in the centuriate and tribal assemblies, senators had to cultivate support among the *populus.* Yet the system was not as democratic as it might appear. Senators with illustrious names and consular ancestors dominated the election to the highest offices, increasing their share of the consulates from about one-half to two-thirds during the 2nd century. These proportions can be interpreted in two ways: the Senate was not a closed, hereditary aristocracy but was open to new families, who usually rose through the senatorial ranks in the course of generations with the patronal support of established families; yet a small circle of prominent families (*e.g.,* the Aemilii, Claudii, and Cornelii) were disproportionately successful, surprisingly so in view of the popular electoral process. Since the campaigning was not oriented toward issues, the great families were able to maintain their superiority over the centuries by their inherited resources: their famous names, their wealth, and their clienteles of voters.

While aristocratic electoral competition was tradition during the republic, this period began to exhibit the escalation in competitiveness that was later fatal to the republic. For example, Publius Cornelius Scipio Africanus emerged from the Second Punic War as the Roman whose *dignitas* (prestige) far surpassed that of his peers. Nonetheless, a number of senators attacked him and his brother Lucius Cornelius with legal charges until he finally retired from Rome to end his life at his Campanian villa at Liternum. For younger senators, however, Scipio's spectacular achievement was something to emulate. The ambitious young Flamininus moved swiftly through the senatorial *cursus honorum* ("course of honors") to win the consulship and command against Philip V at the age of 30. Such cases prompted laws to regulate the senatorial *cursus:* iteration in the same magistracy was prohibited, the praetorship was made a prerequisite for the consulship, and in 180 the *lex Villia annalis* (Villian law on minimum ages) set minimum ages for senatorial magistrates and required a two-year interval between offices. The consulship (two elected to it per year) could be held from age 42, the praetorship (six per year) from age 39, and the curule aedileship from 36. Patricians, still privileged in this area, were probably allowed to stand for these offices two years earlier. The senatorial career was preceded by 10 years of military service, from age 17, and formally began with a quaestorship, the most junior senatorial magistracy (eight per year), at age 30 or just under. The offices between the quaestorship and praetorship, the aedileship (four per year) and the plebeian tribunate (10 per year), were not compulsory but provided opportunities to win popularity among the voters by staging aedilician games and supporting popular causes, respectively. Here again, excess elicited restraint, and legal limits were placed on the lavishness of the games. More broadly, from 181 legislation designed to curb electoral bribery was intermittently introduced.

The problems of electoral competition did not disappear. In the late 150s second consulships were prohibited altogether, but within decades the rules were broken. Scipio Aemilianus, grandson by adoption of Scipio Africanus, challenged the system. Returning from the Carthaginian campaign to Rome to stand for the aedileship, he was elected instead to the consulship, even though he was underage and had not held the prerequisite praetorship. He was then elected to a second consulship for 134. Scipio had no subversive intent, but his career set the precedent for circumventing the *cursus* regulations by appeal to the popular assemblies.

While the 2nd century was a time of heated competition among senators, it was generally a period of quiescence of the plebs and their magistrates, the tribunes. Nevertheless, signs of the upheaval ahead are visible. For one, the long plebeian struggle against arbitrary abuse of magisterial power continued. A series of Porcian laws were passed to protect citizens from summary execution or scourging, asserting the citizen's right of appeal to the assembly (*ius provocationis*). A descendant of the Porcian clan later advertised these laws on coins as a victory for freedom. Moreover, the massive annual war effort provoked occasional resistance to military service. In 193 the tribunes started to investigate complaints about overly long military service. Interpreting this as a challenge to magisterial authority, the Senate responded with a declaration of an emergency levy, and the tribunes stopped their activity. In 151 the tribunes tried to protect some citizens from the levy for the unpopular war in Spain. A confrontation between the tribunes and the recruiting consuls ensued, in which the tribunes briefly imprisoned the consuls until a compromise relieved the crisis. The scene of tribunes taking consuls to jail was repeated in 138 during a period of renewed difficulties over recruiting.

The Porcian laws

Since the Hortensian law of 287, the plebs had the constitutional power to pass laws binding on the entire state without senatorial approval. During the next century and a half few attempts were made to use the power for purposes of major reform against the Senate's will, in part because the plebeian tribunes, as members of the senatorial order, generally shared the Senate's interests and in part because the plebeians benefited from Rome's great successes abroad under senatorial leadership. Yet senatorial fear of unbridled popular legislative power is perceptible in the Aelian and Fufian law of about 150. This law, imperfectly known from later passing references, provided that a magistrate holding a legislative assembly could be prevented from passing a bill on religious grounds by another magistrate claiming to have witnessed unfavourable omens in a procedure called *obnuntiatio.* In addition, the days of the year on which legislative assemblies could be held were reduced. As conservative senators worked to restrain the democratic element in the political processes, the plebeians sought to expand their freedom. Voting in electoral and judicial assemblies had been public, allowing

powerful senators more easily to manage the votes of their clients. The Gabinian law (139) and Cassian law (137) introduced secret written ballots into the assemblies, thus loosening the control of patrons over their clients. Significantly, the reform was supported by Scipio Aemilianus, the sort of senator who stood to benefit by attracting the clients of other patrons through his personal popularity. These reforms, together with the changing composition of the electorate in the city, carried the potential, soon to be realized, for more volatile assemblies.

Culture and religion. Expansion brought Rome into contact with many diverse cultures. The most important of these was the Greek culture in the eastern Mediterranean with its highly refined literature and learning. Rome responded to it with ambivalence: although Greek *doctrina* was attractive, it was also the culture of the defeated and enslaved. Indeed, much Greek culture was brought to Rome in the aftermath of military victories, as Roman soldiers returned home not only with works of art but also with learned Greeks who had been enslaved. Despite the ambivalence, nearly every facet of Roman culture was influenced by the Greeks, and it was a Greco-Roman culture that the Roman empire bequeathed to later European civilization.

As Roman aristocrats encountered Greeks in southern Italy and in the East in the 3rd century, they learned to speak and write in Greek. Scipio Africanus and Flamininus, for example, are known to have corresponded in Greek. By the late republic it became standard for senators to be bilingual. Many were reared from infancy by Greek-speaking slaves and later tutored by Greek slaves or freedmen. Nonetheless, despite their increasing fluency in Greek, senators continued to insist on Latin as the official language of government; visiting dignitaries from the East addressing the Senate in Greek had their speeches translated—as a mark of their subordination.

Because Greek was the lingua franca of the East, Romans had to use Greek if they wished to reach a wider audience. Thus the first histories by Romans were written in Greek. The patrician Fabius Pictor, who, as noted above, founded the Roman tradition of historiography during the Second Punic War, wrote his annalistic history of Rome in Greek partly in order to influence Greek views in favour of Rome, and he emphasized Rome's ancient ties to the Greek world by incorporating in his history the legend that the Trojan hero Aeneas had settled in Latium. Because Roman history was about politics and war, the writing of history was always judged by Romans to be a suitable pastime for men of politics—*i.e.*, for senators such as Fabius.

Rome had had a folk tradition of poetry in the native Saturnian verse with a metre based on stress, but not a formal literature. Lucius Livius Andronicus was regarded as the father of Latin literature, a fact that illustrates to what extent the development of Roman literature was bound up with conquest and enslavement. Livius, a native Greek speaker from Tarentum, was brought as a slave to Rome, where he remained until his death (*c.* 204). Becoming fluent in Latin, he translated the Homeric *Odyssey* into Latin in Saturnian verse. Thus Latin literature began with a translation from Greek into the native metre. Livius reached wider audiences through his translations of Greek plays for public performance. Gnaeus Naevius, the next major figure (*c.* 270–*c.* 201), was again not a native Roman but an Oscan speaker from Campania. In addition to translating Greek drama, he wrote the first major original work in Latin, an epic poem about the First Punic War. Naevius' successors, Quintus Ennius from Calabria (239–169) and Titus Maccius Plautus from Umbria (*c.* 254–184), transformed the Latin poetic genres by importing Greek metrical forms based on the length of syllables rather than on stress. Ennius was best known for his epic history of Rome in verse, the *Annales*, but he also wrote tragedies and satires. Plautus produced comedies adapted from Greek New Comedy. He is the only early author whose work is well represented in the corpus of surviving literature (21 plays judged authentic by Marcus Terentius Varro, Rome's greatest scholar). None of the plays of his younger contemporaries, Caecilius Statius (*c.* 210–168) and Marcus Pacuvius (*c.* 220–130), survive, nor do the once highly esteemed tragedies of Lucius Accius (170–*c.* 86). The six extant comedies of Terence (Publius Terentius Afer; *c.* 190–159) provide a sense of the variation in the comic tradition of the 2nd century. These authors also were outsiders, coming from the Celtic Po valley, Brundisium, Umbria, and North Africa, respectively. Thus, while assorted foreigners, some of servile origin, established a Latin literature by adapting Greek genres, metrical forms, and content, native Roman senators began to write history in Greek.

Literature

Other forms of Greek learning were slower to take root in Rome. Later Romans remembered that a Greek doctor established a practice in Rome for the first time just before the Second Punic War, but his reputation did little to stimulate Roman interest in the subject. Like doctors, Greek philosophers of the 2nd century were regarded with interest and suspicion. In the early 3rd century Romans had erected in public a statue of Pythagoras, a 6th-century Greek philosopher who had founded communities of philosophers in southern Italy. In the mid-2nd century some senators displayed an interest in philosophy. Scipio Aemilianus, Gaius Laelius (consul 140), and Lucius Furius Philus (consul 136) were among those who listened to the lectures of the three leaders of the Athenian philosophical schools visiting Rome on a diplomatic mission in 155—the academic Carneades, the peripatetic Critolaus, and the stoic Diogenes. On an official visit to the East in 140, Scipio included in his entourage the leading stoic Panaetius. In the same period, another stoic, Blossius of Cumae, was said to have influenced the reforming tribune Tiberius Sempronius Gracchus. Yet the philosophical influence should not be exaggerated; none of these senators was a philosopher or even a formal student of philosophy.

Moreover, the sophisticated rhetoric of the philosophers—in 155 Carneades lectured in favour of natural justice one day and against it the next—was perceived by leading Romans such as Cato the Censor as subversive to good morals. At his urging the Senate quickly concluded the diplomatic business of Carneades, Critolaus, and Diogenes in 155 and hurried them out of Rome. This was part of a broader pattern of hostility to philosophy: in 181 the (spurious) Books of Numa, falsely believed to have been influenced by Pythagoras, were burned, and the following decades witnessed several expulsions of philosophers from the city. In comedies of the period, the discipline was held up for ridicule.

Hostility to philosophy

The hostility toward philosophy was one aspect of a wider Roman sense of unease about changing mores. Cato, a "new man" (without senatorial ancestors) elected consul (195) and censor (184), represented himself as an austere champion of the old ways and exemplifies the hardening Roman reaction against change under foreign influence. Although Cato knew Greek and could deploy allusions to Greek literature, he advised his son against too deep a knowledge of the literature of that "most worthless and unteachable race." Cato despised those senatorial colleagues who ineptly imitated Greek manners. He asserted the value of Latin culture in the role of father of Latin prose literature. His treatise on estate management, the *De agricultura* (*c.* 160), has survived with its rambling discourse about how to run a 200-*iugera* (124-acre) farm, including advice on everything from buying and selling slaves to folk medicine. Cato's greater, historical work, the *Origines*, survives only in fragments: it challenged the earlier Roman histories insofar as it was written in Latin and emphasized the achievements of the Italian peoples rather than those of the few great senatorial families of Rome (whose names were conspicuously omitted).

Elected censor in 184 to protect Roman mores, Cato vowed "to cut into pieces and burn like a hydra all luxury and voluptuousness." He expelled seven men from the Senate on various charges of immorality and penalized through taxation the acquisition of such luxuries as expensive clothing, jewelry, carriages, and fancy slaves. The worry about luxury was widespread, as evidenced by the passage of a series of sumptuary laws supported by Cato. During the depths of the Second Punic War the Oppian law (215) was passed to meet the financial crisis by re-

stricting the jewelry and clothing women were allowed to wear; in 195, after the crisis, the law was repealed despite Cato's protests. Later sumptuary laws were motivated not by military crisis but by a sense of the dangers of luxury: the Orchian law (182) limited the lavishness of banquets; the Fannian law (161) strengthened the Orchian provisions, and the Didian law (143) extended the limits to all Italy. A similar sense of the dangers of wealth may also have prompted the *lex Voconia* (169), which prohibited Romans of the wealthiest class from naming women as heirs in their wills.

The laws and censorial actions ultimately could not restrain changes in Roman mores. Economic conditions had been irreversibly altered by conquest; the magnitude of conspicuous consumption is suggested by a senatorial decree of 161 that restricted the weight of silver tableware in a banquet to 100 pounds—10 times the weight for which Publius Cornelius Rufinus was punished in 275. Moreover, the very competitiveness that had traditionally marked the senatorial aristocracy ensured the spread of cultural innovations and new forms of conspicuous consumption among the elite. In contrast to the austere Cato, other senators laid claim to prestige by collecting Greek art and books brought back by conquering armies, by staging plays modeled on Greek drama, and by commissioning literary works, public buildings, and private sculptural monuments in a Greek style.

Whereas the influence of Greek high culture was felt principally in a small circle of elite Romans who had the wealth to acquire Greek art and slaves and the leisure and education to read Greek authors, the influence of religions from the eastern Mediterranean was perceived as potentially subversive to a far wider audience. Polybius praised the Romans for their conscientious behaviour toward the gods. Romans were famous for their extreme precision in recitation of vows and performance of sacrifices to the gods, meticulously repeating archaic words and actions centuries after their original meanings had been forgotten. Guiding these state cults were priestly colleges; and priestly offices such as of pontifex and augur were filled by senators, whose dominance in politics was thus replicated in civic religion.

In earlier centuries Rome's innate religious conservatism was, however, counterbalanced by an openness to foreign gods and cults. As Rome incorporated new peoples of Italy into its citizen body, it accepted their gods and religious practices. Indeed, among the most authoritative religious texts, consulted in times of crisis or doubt, were the prophetic Sibylline Books, written in Greek and imported from Cumae. The receptivity appears most pronounced in the 3rd century: during its final decades temples were built in the city for Venus Erycina from Sicily and for the Magna Mater, or Great Mother, from Pessinus in Anatolia; games were instituted in honour of the Greek god Apollo (212) and the Magna Mater after the war. The new cults were integrated into the traditional structure of the state religion, and the "foreignness" was controlled (*i.e.*, limits were placed on the orgiastic elements in the cult of the Great Mother performed by her eunuch priests).

Openness to foreign cults

The openness, never complete or a matter of principle, tilted toward resistance in the early 2nd century. In 186 Roman magistrates, on orders from the Senate, brutally suppressed Bacchic worship in Italy. Associations of worshipers of the Greek god Bacchus (Dionysus) had spread across Italy to Rome. Their members, numbering in the thousands, were initiated into secret mysteries, knowledge of which promised life after death; they also engaged in orgiastic worship. The secrecy soon gave rise to reports of the basest activities, such as uncontrolled drinking, sexual promiscuity, forgery of wills, and poisoning of kin. According to Livy, more than 7,000 were implicated in the wrongdoing; many of them were tried and executed, and the consuls destroyed the places of Bacchic worship throughout Italy. For the future, the (extant) senatorial decree prohibited men from acting as priests in the cult, banned secret meetings, and required the praetor's and Senate's authorization of ceremonies to be performed by gatherings of more than five people. The terms of the decree provide a sense of what provoked the harsh senatorial reaction. It was not that the Bacchic cult spread heretical beliefs about the gods—Roman civic religion was never based on theological doctrine with pretensions to exclusive truth; rather, the growing secret cult led by male priests threatened the traditionally dominant position of senators in state religion. The decree did not aim to eliminate Bacchic worship but to bring it under the supervision of senatorial authorities. The following centuries witnessed sporadic official actions against foreign cults; it happens to be recorded that a praetor of 139 removed private altars built in public areas and expelled astrologers and Jews from the city. Thus the reaction to eastern religions paralleled that to Greek philosophy; both were perceived as new ways of thinking that threatened to undermine traditional mores and the relations of authority implicit in them.

Economy and society. It seems certain that the economy and society of Italy were transformed in the wake of Rome's conquest of the Mediterranean world, even though the changes can be described only incompletely and imprecisely, owing to the dearth of reliable information for the preceding centuries. Romans of the 1st century BC believed that their ancestors had been a people of small farmers in an age uncorrupted by wealth. Even senators who performed heroic feats were said to have been of modest means—men such as Lucius Quinctius Cincinatus, who was said to have laid down his plow on his tiny farm to serve as dictator in 458 BC. Although such legends present an idealized vision of early Rome, it is probably true that Latium of the 5th and 4th centuries was densely populated by farmers of small plots. Rome's military strength derived from its superior resources of manpower levied from a pool of small landowning citizens (*assidui*). A dense population is also suggested by the emigration from Latium of scores of thousands as colonists during the 4th and 3rd centuries. The legends of senators working their own fields seem implausible, but the disparity in wealth was probably much less noticeable than in the late republic. The 4th-century artifacts uncovered by archaeologists display an overall high quality that makes it difficult to distinguish a category of luxury goods from the pottery and terra-cottas made for common use.

A people of small farmers

War and conquest altered this picture; yet certain fundamental features of the economy remained constant. Until its fall, the Roman Empire retained agriculture as the basis of its economy, with probably four-fifths of the population tilling the soil. This great majority continued to be needed in food production because there were no labour-saving technological breakthroughs. The power driving agricultural and other production was almost entirely supplied by humans and animals, which set modest limits to economic growth. In some areas of Italy, such as the territory of Capena in southern Etruria, archaeologists have found traditional patterns of settlement and land division continuing from the 4th to the end of the 1st century—evidence that the Second Punic War and the following decades did not bring a complete break with the past.

Economic change came as a result of massive population shifts and the social reorganization of labour rather than technological improvement. The Second Punic War, and especially Hannibal's persistent presence in Italy, inflicted a considerable toll, including loss of life on a staggering scale, movement of rural populations into towns, and destruction of agriculture in some regions. Although the devastation has been overestimated by some historians, partial depopulation of the Italian countryside is evident from the literary and archaeological records: immediately after the war enough land stood vacant in Apulia and Samnium to settle between 30,000 and 40,000 of Scipio's veterans, while areas of Apulia, Bruttium, southern Campania, and south-central Etruria have yielded no artifacts indicating settlement in the postwar period.

Populations have been known to show great resilience in recovering from wars, but the Italian population was given no peace after 201. In subsequent decades Rome's annual war effort required a military mobilization unmatched in history for its duration and the proportion of the population involved. During the 150 years after Hannibal's surrender, the Romans regularly fielded armies of more than 100,000 men, requiring on average about 13

percent of the adult male citizens each year. The attested casualties from 200 to 150 add up to nearly 100,000. The levy took Roman peasants away from their land. Many never returned. Others, perhaps 25,000, were moved in the years before 173 from peninsular Italy to the colonies of the Po valley. Still others, in unknown but considerable numbers, migrated to the cities. By the later 2nd century some Roman leaders perceived the countryside to be depopulated.

Slavery

To replace the peasants on the land of central and southern Italy, slaves were imported in vast numbers. Slavery was well established as a form of agricultural labour before the Punic Wars (slaves must have produced much of the food during the peak mobilization of citizens from 218 to 201). The scale of slavery, however, increased in the 2nd and 1st centuries as a result of conquests. Enslavement was a common fate for the defeated in ancient warfare: the Romans enslaved 5,000 Macedonians in 197; 5,000 Histri in 177; 150,000 Epirotes in 167; 50,000 Carthaginians in 146; and in 174 an unspecified number of Sardinians, but so many that "Sardinian" became a byword for "cheap" slave. These are only a few examples for which the sources happen to give numbers. More slaves flooded into Italy after Rome destabilized the eastern Mediterranean in 167 and gave pirates and bandits the opportunity to carry off local peoples of Anatolia and sell them on the block at Delos by the thousands. By the end of the republic Italy was a thoroughgoing slave society with well over one million slaves, according to the best estimates. No census figures give numbers of slaves, but slaveholding was more widespread and on a larger scale than in the antebellum American South, where slaves made up about one-third of the population. In effect, Roman soldiers fought in order to capture their own replacements on the land in Italy, although the shift from free to servile labour was only a partial one.

The influx of slaves was accompanied by changes in patterns of landownership, as more Italian land came to be concentrated in fewer hands. One of the punishments meted out to disloyal allies after the Second Punic War was confiscation of all or part of their territories. Most of the *ager Campanus* and part of the Tarentines' lands—perhaps two million acres in total—became Roman *ager publicus* (public land), subject to rent. Some of this property remained in the hands of local peoples, but large tracts in excess of the 500-*iugera* limit were occupied by wealthy Romans, who were legally *possessores* (*i.e.*, in possession of the land, although not its owners) and as such paid a nominal rent to the Roman state. The trend toward concentration continued during the 2nd century, propelled by conquests abroad. On the one side, subsistance farmers were always vulnerable in years of poor harvests that could lead to debt and ultimately to the loss of their plots. The vulnerability was exacerbated by army service, which took peasants away from their farms for years at a time. On the other side, the elite orders were enriched by the booty from the eastern kingdoms on a scale previously unimaginable. Some of the vast new wealth was spent on public works and on new forms of luxury and part was invested to secure future income. Land was the preferred form of investment for senators and other honourable men: farming was regarded as safer and more prestigious than manufacture or trade. For senators, the opportunities for trade were limited by the Claudian law of 218 prohibiting them from owning large ships. Wealthy Romans thus used the proceeds of war to buy out their smaller neighbours. As a result of this process of acquisition, most senatorial estates consisted of scattered small farms. The notorious *latifundia*, the extensive consolidated estates, were not widespread. Given the dispersion of the property, the new landlord was typically absentee. He could leave the working of the farms in the hands of the previous peasant owners as tenants, or he could import slaves.

Cato's *De agricultura*

The best insights into the mentality of the estate-owning class of this period come from Cato's *De agricultura*. Although based on Greek handbooks discussing estate management, it reflects the assumptions and thinking of a 2nd-century senator. Cato envisaged a medium-sized, 200-*iugera* farm with a permanent staff of 11 slaves. As with other Roman enterprises, management of the farm was left to a slave bailiff, who was helped by his slave wife. While Cato, like the later agricultural writers Varro and Lucius Junius Columella, assumed the economic advantage of a slave work force, historians today debate whether estates worked by slaves were indeed more profitable than smaller peasant farms. Cato had his slaves use much the same technology as the peasants, although a larger estate could afford large processing implements, such as grape and olive crushers, which peasants might have to share or do without. Nor did Cato bring to bear any innovative management advice; his suggestions aimed to maximize profits by such commonsense means as keeping the slave work force occupied all year round and buying cheap and selling dear. Nevertheless, larger estates had one significant advantage in that the slave labour could be bought and sold and thus more easily matched to labour needs than was possible on small plots worked by peasant families.

Cato's farm was a model representing one aspect of the reality of the Italian countryside. Archaeologists have discovered the villas characteristic of the Catonian estate beginning to appear in Campania in the 2nd century and later in other areas. The emergence of slave agriculture did not exclude the continuing existence in the area of peasants as owners of marginal land or as casual day labourers or both. The larger estates and the remaining peasants formed a symbiotic relationship, mentioned by Cato: the estate required extra hands to help during peak seasons, while the peasants needed the extra wages from day labour to supplement the meagre production of their plots. Yet in many areas of Italy the villa system made no inroads during the republic, and traditional peasant farming continued. Other areas, however, underwent a drastic change: the desolation left by the Second Punic War in the central and southern regions opened the way for wealthy Romans to acquire vast tracts of depopulated land to convert to grazing. This form of extensive agriculture produced cattle, sheep, and goats, herded by slaves. These were the true *latifundia*, decried as wastelands by Roman imperial authors such as the elder Pliny.

The marketplace took on a new importance as both the Catonian estate and the *latifundium* aimed primarily to produce goods to sell for a profit. In this sense, they represented a change from peasant agriculture, which aimed above all to feed the peasant's family. The buyers of the new commodities were the growing cities—another facet of the complex economic transformation. Rome was swelled by migrants from the countryside and became the largest city of preindustrial Europe, with a population of about one million in the imperial era; other Italian cities grew to a lesser extent.

The mass of consumers created new, more diverse demands for foodstuffs from the countryside and also for manufactured goods. The market was bipolar, with the poor of the cities able to buy only basic foodstuffs and a few plain manufactured items and the rich demanding increasingly extravagant luxury goods. The limitations of the poor are reflected in the declining quality of humble temple offerings. The craftsmen and traders produced mainly for the rich minority. The trading and artisanal enterprises in Rome were largely worked by slaves and freedmen imported to Rome by the wealthy. Although honourable, freeborn Romans considered it beneath their dignity to participate directly in these businesses, they willingly shared in the profits through ownership of these slaves and through collection of rents on the shops of humbler men. Thus, manufacturing and trading were generally small-scale operations, organized on the basis of household or family. Roman law did not recognize business corporations with the exception of publican companies holding state contracts; nor were there guilds of the medieval type to organize or control production. Unlike some later medieval cities, Rome did not produce for export to support itself; its revenues came from booty, provincial taxes, and the surplus brought from the countryside to the city by aristocratic Roman landlords. Indeed, after 167 provincial revenues were sufficient to allow for the abolition of direct taxes on Roman citizens.

Trading and artisanal enterprises

Building projects were the largest enterprises in Rome

and offered freeborn immigrants jobs as day labourers. In addition to the private building needed to house the growing population, the early and middle 2nd century witnessed public building on a new scale and in new shapes. The leading senatorial families gained publicity by sponsoring major new buildings named after themselves in the Forum and elsewhere. The Basilica Porcia (built during Marcus Porcius Cato's censorship of 184), the Basilica Aemilia et Fulvia (179), and the Basilica Sempronia (170–169) were constructed out of the traditional tufa blocks but in a Hellenized style.

New infrastructures were required to bring the necessities of life to the growing population. The Porticus Aemilia (193), a warehouse of 300,000 square feet on the banks of the Tiber, illustrates how the new needs were met with a major new building technology, concrete construction. Around 200 BC in central Italy it was discovered that a wet mixture of crushed stone, lime, and sand (especially a volcanic sand called *pozzolana*) would set into a material of great strength. This construction technique had great advantages of economy and flexibility over the traditional cut-stone technique: the materials were more readily available, the concrete could be molded into desired shapes, and the molds could be reused for repetitive production. The Porticus Aemilia, for example, consisted of a series of roughly identical arches and vaults—the shapes so characteristic of later Roman architecture. The new technology also permitted improvements in the construction of the aqueducts needed to increase the city's water supply.

The economic development outside of Rome encompassed some fairly large-scale manufacturing enterprises and export trade. At Puteoli on the Bay of Naples the ironworks industry was organized on a scale well beyond that of the household, and its goods were shipped beyond the area. Puteoli flourished during the republic as a port city, handling imports destined for Rome as well as exports of manufactured goods and processed agricultural products. In their search for markets, the large Italian landowners exported wine and olive oil to Cisalpine Gaul and more distant locations. Dressel I amphoras, the three-foot pottery jars carrying these products, have been found in substantial quantities in Africa and Gaul. Yet the magnitude of the economic development should not be exaggerated: the ironworks industry was exceptional, and most pottery production continued to be for local use.

Social changes

Major social changes and dislocations accompanied the demographic shifts and economic development. Relations between rich and poor in Rome had traditionally been structured by the bond existing between patron and client. In the daily morning ritual of the *salutatio,* humble Romans went to pay their respects in the houses of senators, who were obligated to protect them. These personal relationships lent stability to the social hierarchy. In the 2nd century, however, the disparity between rich and poor citizens grew. While this trend increased the personal power of individual senators, it weakened the social control of the elite as a whole; the poor had become too numerous to be controlled by the traditional bond of patron and client.

Until the end of the 170s the impoverishment of humble citizens had been counterbalanced to some extent by the founding of colonies, because dispossessed peasants were given new lands in outlying regions. During the middle decades of the 2nd century, however, colonization ceased, and the number of dispossessed increased, to judge from the declining number of small landowners in the census. The problem created by a growing proletariat was recognized by a few senators. Gaius Laelius, probably during his consulship of 140, proposed a scheme of land redistribution to renew the class of smallholders, but it was rejected by the Senate.

Some of the dispossessed went to Rome, where, together with the increasing numbers of slaves and freedmen, they contributed to the steadily growing population. This density led to the miseries associated with big cities, which were exacerbated by the absence of regulation. By 200 BC the pressure of numbers necessitated apartment buildings of three stories. Constructed without a building code, these structures were often unsound and prone to collapse. Moreover, closely placed and partly made of wood, they were tinderboxes, ever ready to burst into flame. The population density also increased the vulnerability to food shortages and plagues. In 188 fines were levied against dealers for withholding grain, attesting to problems of supply. The 180s and 170s witnessed repeated outbreaks of plague. The state, which could use its power to increase the grain supply, was helpless against diseases. In general, the republican state developed few new institutions to manage the growing urban problems: until the reign of Augustus matters were left to the traditional authority of urban magistrates, who were unaided by a standing fire brigade or police force. Consequently, Rome held an increasing potential for social discontent and conflicts without a corresponding increase in means of control.

The family, regarded by Romans as a mainstay of the social order, also was affected by the wider economic and social transformations of the 2nd century BC. In the early republic the family had formed a social, economic, and legal unity. The woman generally married into her husband's family and came under his legal authority (or that of his father if he was still alive), and her dowry merged with the rest of the estate under the ownership of the husband. The husband managed the family's affairs outside the house, while the wife was custodian within. Marriage was an arrangement for life; divorces were rare and granted only in cases of serious moral infractions, such as adultery or wine-tippling on the part of the wife. The children of the couple were subject to the father's nearly absolute legal powers (*patria potestas*), including the power of life and death, corporal punishment, and a monopoly of ownership of all property in the family. The father's power lasted until his death or, in the case of a daughter, until her marriage. When the father died, his sons, his wife, and his unmarried daughters became legally independent, and all inherited equal shares of the family's property unless otherwise specified in a will. The imperial authors idealized the early republic as a time of family harmony and stability, which was lost through the corruption of the later republic.

"Free marriage"

When family life emerged into the full light of history in the 2nd century BC, it had changed in significant ways. A form of marriage, commonly called "free marriage," was becoming prevalent. Under this form, the wife no longer came into her husband's power or property regime but remained in that of her father; upon her father's death she became independent with rights to own and dispose of property. But she was not a member of the family of her husband and children and had no claim to inheritance from them, even though she lived with them in the same house. Because many women inherited part of their fathers' estates, they could use their independent fortunes to exert influence on husbands, children, and people outside the house. In the same period divorce became far more common; moral infractions were no longer needed to justify divorce, which could be initiated by either side. Frequent divorce and remarriage went hand in hand with the separation of marital property. There is plausibility in the suggestion that these changes were brought on by a desire of the women's fathers to avoid having their daughters' portions of the larger family estates slip irrevocably into the hands of their husbands. Although the changes in law and practice were not motivated by any movement to emancipate women, the result was that propertied women of the late republic, always excluded from the public sphere of male citizens, came to enjoy a degree of freedom and social power unusual before the 20th century.

Slaves came to permeate the fabric of family life and altered relationships within the household. They were regularly assigned the tasks of child-rearing, traditionally the domain of the mother, and of education, until then the responsibility of both the father and the mother. Whereas children had acquired the skills needed for their future roles by observing their parents in a kind of apprenticeship, in wealthy houses sons and, to a lesser extent, daughters were now given a specialized education by slaves or freedmen. The management of aristocratic households was entrusted to slaves and freedmen, who served as secretaries, accountants, and managers. The wife was no longer needed as custodian of the household, though domestic

guardianship remained an element in the idealization of her role. Later moralists attributed a decline in Roman virtue and discipline to the intrusion of slaves into familial relationships and duties.

Rome and Italy. During the middle republic the peoples of Italy began to coalesce into a fairly homogeneous and cohesive society. Polybius, however, does not give insight into this process, because, living in Rome, he too little appreciated the variety of Italian cultures under Roman sway, from the Gallic peoples in the mountains of the north to the urbane Greeks on the southern coasts. Other evidence, though meagre, nonetheless suggests several processes that contributed to the increasing cohesion.

First, the Romans built a network of roads that facilitated communication across Italy. As stated above, the first great road was the Via Appia, which was laid out by Appius Claudius Caecus in 312 to connect Rome to Capua. Between the First and Second Punic Wars roads were built to the north: the Via Aurelia (241?) along the Tyrrhenian coast, the Via Flaminia (220) through Umbria, and the Via Clodia through Etruria. Then, in the 2nd century, Roman presence in the Po valley was consolidated by the Via Aemilia (187) from Ariminum on the Adriatic coast to the Latin colony of Placentia and by the Via Postumia (148) running through Transpadane Gaul to Aquileia in the east and Genua in the west.

Second, internal migration—Italians moving to Rome and Romans being sent to Latin colonies throughout Italy—promoted social and cultural homogeneity. Some of these colonies were set alongside existing settlements; others were founded on new sites. The colonies re-created the physical and social shape of Rome; the town plans and architecture, with forums including temples to Jupiter, were modeled on those of Rome. The imposition of a Latin colony on the Greek city of Paestum in Lucania (273) entailed the implantation of a Roman-style forum in the centre of the existing city in a way that rudely intruded on the old sanctuary of Hera. The initial system governing the distribution of land to Latin colonists aimed to replicate the Roman social hierarchy differentiated by wealth: it is recorded of the colonists sent to Aquileia in 181 that the 3,000 infantrymen each received 50 *iugera* (31 acres), the centurions 100 *iugera* (62 acres), and the cavalrymen 140 *iugera* (86 acres). The unifying effect of the colonies is evident in Paestum's notable loyalty to Rome during the Second Punic War.

Latin colonies in Italy

Third, although Rome did not seek to govern Italy through a regular administration, it influenced local affairs through formal bonds of personal friendship (*amicitia*) and hospitality (*hospitium*) between the Roman elite and their local counterparts. Through these ties the leading men of Italy were gradually drawn into the ruling class in Rome. The most prominent example of the 2nd century is that of Gaius Marius of Arpinum, who, only two generations after his town had received full citizen rights, began his meteoric senatorial career under the patronage of the great Roman nobles, the Metelli.

Fourth, the regular military campaigns brought together Romans and Italians of all classes under the command of Roman magistrates. The Italian troops appear to have been levied in a fashion similar to the one used for the Romans, which would have required a Roman-style census as a means of organizing the local citizenries. In the absence of direct administration, military service was the context in which Italians most regularly experienced Roman authority.

Fifth, Rome occasionally deployed its troops in Italy to maintain social order. Rome suppressed an uprising of serfs in Etruscan Volsinii in 265 and a sedition in Patavium in 175. When the massive influx of slaves raised the spectre of rebellions across Italy, Roman troops were deployed to put down uprisings: in 195, 5,000 slaves were executed in Latin Setia; in 196 the praetor was sent with his urban legion to Etruria to fight a pitched battle in which many slaves were killed; and the praetor of 185 dealt with rebellious slaves in Apulia, condemning 7,000 to death. The later slave revolt in Sicily (*c.* 135–132) was not contained so effectively and grew to include perhaps 70,000. The slaves defeated the first consular army sent in 134; the efforts of two more consuls were required to restore order. The revolts, unusual for their frequency and size, are not to be explained by abolitionist programs (nonexistent in antiquity) nor by maltreatment. The causes lay in the enslavement and importation of entire communities with their native leadership and in the free reign given to slave shepherds who roamed armed around the countryside serving as communication lines between slave plantations. These uprisings made it clear that the social fabric of Italy, put under stress by the transformations brought about by conquest, had to be protected by Roman force.

While the exercise of Roman authority and force was sometimes resented by Italians, Rome's power made its mores and culture worthy of imitation. The Latin language and Roman political institutions slowly spread. A request from the old Campanian city of Cumae in 180 that it be allowed to change its official language from Oscan to Latin was a sign of things to come. (R.P.Sa.)

THE LATE REPUBLIC (133–31 BC)

The aftermath of the victories. The fall of Carthage and Corinth did not even mark a temporary end to warfare. War and military glory were an essential part of the Roman aristocratic ethos and, hence, of Roman political life. Apart from major wars still to come, small wars on the frontiers of Roman power—never precisely fixed—continued to provide an essential motive in Roman history: in Spain, Sardinia, Illyria, and Macedonia, barbarians could be defeated and triumphs won. Thus the limits of Roman power were gradually extended and the territories within them pacified, while men of noble stock rivaled the *virtus* of their ancestors and new men staked their own competing claims, winning glory essential to political advancement and sharing the booty with their officers and soldiers. Cicero could still depict it as a major disgrace for Lucius Piso (consul; 58 BC) that he had won no triumph in the traditionally "triumphal" province of Macedonia. Nonetheless, the coincidence of the capture of Corinth and Carthage was even in antiquity regarded as a turning point in Roman history: it was the end (for the time being) of warfare against civilized powers, in which the danger was felt to be greater and the glory and the booty were superior to those won against barbarian tribes.

Changes in provincial administration. The first immediate effect was on the administration of the empire. The military basis of provincial administration remained: the governor (as he is called) was in Roman eyes a commander with absolute and unappealable powers over all except Roman citizens, within the limits of the territory (his *provincia*) assigned to him (normally) by the Senate. He was always prepared—and in some provinces expected—to fight and win. But it had been found that those unlimited powers were often abused and that Senate control could not easily be asserted at increasing distances from Rome. For political and perhaps for moral reasons, excessive abuse without hope of a remedy could not be permitted. Hence, when the decision to annex Carthage and Macedonia had been made in principle (149 BC), a permanent court (the *quaestio repetundarum*) was established at Rome to hear complaints against former commanders and, where necessary, to assure repayment of illegal exactions. No penalty for offenders was provided, and there was no derogation from the commander's powers during his tenure; nevertheless, the step was a landmark in the recognition of imperial responsibility, and it was also to have important effects on Roman politics.

Recognition of imperial responsibility

Another result of the new conquests was a major administrative departure. When Africa and Macedonia became *provinciae* to be regularly assigned to commanders, it was decided to break with precedent by not increasing the number of senior magistrates (praetors). Instead, prorogation—the device of leaving a magistrate in office *pro magistratu* ("in place of a magistrate") after his term had expired, which had hitherto been freely used when emergencies had led to shortages of regular commanders—was established as part of the administrative system: thenceforth, every year at least two praetors would have to be retained as promagistrates. This was the beginning of the

dissociation between urban magistracy and foreign command that was to become a cardinal principle of the system of Sulla and of the developed Roman Empire.

Social and economic ills. It is not clear to what extent the temporary end of the age of major wars helped to produce the crisis of the Roman Republic. The general view of thinking Romans was that the relaxation of external pressures led to internal disintegration. (This has happened in other states, and the view is not to be lightly dismissed.) Moreover, the end of large-scale booty led to economic recession in Rome, thus intensifying poverty and discontent. But the underlying crisis had been building up over a long period.

The reform movement of the Gracchi (133–121 BC). From the state's point of view, the chief effect was a decline in military manpower. The minimum property qualification for service was lowered and the minimum age (17) ignored; resistance became frequent, especially to the distant and unending guerrilla war in Spain.

The program and career of Tiberius Sempronius Gracchus. Tiberius Gracchus, grandson of Scipio Africanus and son of the Gracchus who had conquered the Celtiberi and treated them well, was quaestor in Mancinus' army when it faced annihilation; on the strength of his family name, he personally negotiated the peace that saved it. When the Senate—on the motion of his cousin Scipio Aemilianus, who later finished the war—renounced the peace, Tiberius felt aggrieved; he joined a group of senior senators hostile to Aemilianus and with ideas on reform. Elected tribune for 133, in Scipio's absence, Tiberius attempted to find a solution for the social and military crisis, with the political credit to go to himself and his backers. Tiberius had no intention of touching private property; his idea was to enforce the legal but widely ignored limit of 500 *iugera* (309 acres) on occupation of public land and to use the land thus retrieved for settling landless citizens, who would both regain a secure living and be liable for service. The slave war in Sicily, which had lasted several years and had threatened to spread to Italy, had underlined both the danger of using large numbers of slaves on the land and the need for a major increase in military citizen manpower.

Plan to settle landless citizens

Tiberius' proposal was bound to meet with opposition in the Senate, which consisted of large landowners. On the advice of his eminent backers, he took his bill—which made various concessions to those asked to obey the law and hand back excess public land—straight to the Assembly of the Plebs, where it found wide support. This procedure was not revolutionary; bills directly concerning the people appear to have been frequently passed in this way. But his opponents persuaded another aristocratic tribune, Marcus Octavius, to veto the bill. Tiberius tried the constitutional riposte: an appeal to the Senate for arbitration. But the Senate was unwilling to help, and Octavius was unwilling to negotiate over his veto—an action apparently unprecedented, though not (strictly speaking) unconstitutional. Tiberius had to improvise a way out of the impasse. He met Octavius' action with a similarly unprecedented retort and had Octavius deposed by the Assembly. He then passed his bill in a less conciliatory form and had himself, his father-in-law, and his brother appointed commissioners with powers to determine boundaries of public land, confiscate excess acreage, and divide it in inalienable allotments among landless citizens. As it happened, envoys from Pergamum had arrived to inform the Senate that Attalus III had died and made the Roman people his heirs (provided the cities of his kingdom were left free). Tiberius, at whose house the envoys were lodging, anticipated Senate debate and had the inheritance accepted by the people and the money used to finance his agrarian schemes.

Tiberius' opponents now charged him with aiming at tyranny, a charge that many may well have believed: redistribution of land was connected with demagogic tyranny in Hellenistic states, and Tiberius' subsequent actions had been high-handed and beyond the flexible borderline of what was regarded as *mos majorum* (constitutional custom). Fearing prosecution once his term in office was over, he now began to canvass for a second tribunate—another unprecedented act, bound to reinforce fears of tyranny. The elections took place in an atmosphere of violence, with nearly all his tribunician colleagues now opposed to him. When the consul Publius Scaevola, on strict legal grounds, refused to act against him, Publius Scipio Nasica, the chief ἀνοντίῳ, led a number of senators and their clients to the Assembly, and Tiberius was killed in a resulting scuffle. Widespread and bloody repression followed in 132. Thus political murder and political martyrdom were introduced into Roman politics.

The land commission

The land commission, however, was allowed to continue because it could not easily be stopped. Some evidence of its activities survives. By 129, perhaps running out of available land held by citizens, it began to apply the Gracchan law to public land held by Italian individuals or communities. This had probably not been envisaged by Tiberius, just as he did not include noncitizens among the beneficiaries of distributions. The Senate, on the motion of Scipio Aemilianus, upheld the Italians' protests, transferring decisions concerning Italian-held land from the commission to a consul. This seriously hampered the commission's activities. Marcus Fulvius Flaccus, chairman of the commission and consul in 125, tried to solve the problem by offering the Italians the citizenship (or alternatively the right to appeal against Roman executive acts to the Roman people) in return for bringing their holdings of public land under the Gracchan law. This aroused fears of uncontrollable political repercussions. Flaccus was ordered by the Senate to fight a war in southern France (where he gained a triumph) and had to abandon his proposal. There is no sign of widespread Italian interest in it at this time, though the revolt of the Latin colony Fregellae (destroyed 125) may be connected with its failure.

The program and career of Gaius Sempronius Gracchus. In 123 Gaius Gracchus, a younger brother of Tiberius, became tribune. He had served on Tiberius' land commission and had supported Flaccus' plan. Making the most of his martyred brother's name, Gaius embarked on a scheme of general reform in which, for the first time in Rome, Greek theoretical influences may be traced. Among many reforms—including provision for a stable and cheap wheat price and for the foundation of colonies (one on the site of Carthage), to which Italians were admitted—two major ideas stand out: to increase public revenues (both from the empire and from taxes) and pass the benefit on to the people; and to raise the wealthiest nonsenators (particularly the equites, holders of the "public horse"—who received state financial aid for the purchase and upkeep of their horses—and next to senators in social standing) to a position from which, without actually taking part in the process of government, they could watch over senatorial administration and make it more responsible. The idea was evoked by Tiberius' death. As early as 129 a law compelled senators to surrender the "public horse" (which hitherto they had also held) and possibly in other ways enhanced the group consciousness and privileges of the equites. Regarding the increase of public revenue, Gaius put the *publicani* (public contractors, hitherto chiefly concerned with army and building contracts and with farming minor taxes) in charge of the main tax of Asia—a rich province formed out of Attalus' inheritance, which would henceforth provide Rome with the major part of its income. This was expected both to reduce senatorial corruption and to improve efficiency. Gaius also put eminent nonsenators (probably defined by wealth, but perhaps limited to the equites, or equestrian class) in charge of the *quaestio repetundarum*, whose senatorial members had shown too much leniency to their colleagues, and he imposed severe penalties on senators convicted by that court. Finally, in a second tribunate, he hoped to give citizenship to Latins and Latin rights to other Italians, with the help of Flaccus who, though a distinguished former consul, took the unique step of becoming tribune. But a consul and a tribune of 122 together persuaded the citizen voters that it was against their interests to share the privileges of citizenship: the bill was defeated, and Gaius failed in his attempt to be re-elected once more. In 121, preparing (as private citizens) to use force to oppose the cancellation of some of their laws, Gaius and Flaccus were killed in a riot, and many of their followers were executed.


VICINITY OF ROME
0 5 10 mi
Main roads
Branch roads
Aqueducts
1 Aqua Appia
2 Aqua Tripula
3 Aqua Virgo
4 Forentinian Grove
5 Hadrian's Villa
6 Milvian Bridge
7 Temple of Jupiter Latiaris
Port of Augustus
Port of Claudius
Port of Trajan
Ostia
Rome
Tivoli
Tusculum
Praeneste
Lavinium
Ardea
Norbe
Sigma
Cora
Pedum
Aefulae
Subiaco
Tibur
Gabii
Lucretilis
Mt. Albanus
Alba Longa
Aricia
Nemorensis
Lake Nemi
Velitrae
Anagnia
Arpinum
Fregenae
Lorium
Veii
Caere
Nepet
Sutrium
Falerii
Capena
Cures
Eretum
Cameria
Nomentum
Fidenae
Collatia
Gabii
Praeneste
Labicum
Lucullum
Tellenae
Carsioli
Alba Fucens
Marsi
Sabina
Forum Clodii
Lake Sabatinus
Ad Novas
Baccanae
Arsiolum
Veii
Caere
Nepet
Sutrium
Falerii
Capena
Cures
Eretum
Cameria
Nomentum
Fidenae
Collatia
Gabii
Praeneste
Labicum
Lucullum
Tellenae
Carsioli
Alba Fucens
Marsi
Sabina
Forum Clodii
Lake Sabatinus
Ad Novas
Baccanae
Arsiolum
Via Flaminia
Via Cassia
Via Clodia
Via Aurelia
Via Tiburtina
Via Nomentana
Via Salaria
Via Labicana
Via Praenestina
Via Latina
Via Appia
Via Ardeatina
Via Ostiensis
Via Portuensis
Via Triumphalis
Via Campana
Via Valeria
Via Sublacensis
Via Nova
Via Gallica
Via Tiburtina
Via Collatina
Via Pr

During the next decade the measures benefiting the people were largely abolished, though the Gracchan land distributions, converted into private property, did temporarily strengthen the Roman citizen peasantry. The provisions giving power to wealthy nonsenators could not be touched, for political reasons, and they survived as the chief effect of Gaius' tribunates. The court seems to have worked better than before, and, during the next generation, several other standing criminal courts were instituted, as were occasional ad hoc tribunals, always with the same class of jurors. In 106 a law adding senators to the juries was passed, but it remained in force for only a short time.

The republic (c. 121–91 BC). *War against Jugurtha.* Since Roman historians were no more interested in politics than (on the whole) in social or economic developments, the struggles of the aristocratic families must be pieced together from chance information. It would be mere paradox to deny the importance in republican Rome, as in better known aristocratic republics, of family feuds, alliances, and policies, and parts of the picture are known—*e.g.,* the central importance of the family of the Metelli, prominent in politics for a generation after the Gracchi and dominant for part of that time. In foreign affairs the client kingdom of Numidia—loyal ever since its institution by Scipio Africanus—assumed quite unwarranted importance when a succession crisis developed there soon after 120, as a bastard, Jugurtha, relying on superior ability and aristocratic Roman connections, sought to oust his two legitimate brothers from their shares of the divided kingdom. Rome's usual diplomatic methods failed to stop Jugurtha from disposing of his brothers, but the massacre of Italian settlers at Cirta by his soldiers forced the Senate to declare war (112). The war was waged reluctantly and ineffectively, with the result that charges of bribery were freely bandied about by demagogic tribunes taking advantage of suspicion of aristocratic political behaviour that had smoldered ever since the Gracchan crisis. Significantly, some eminent men, hated from those days, were now convicted of corruption. The Metelli, however, emerged unscathed, and Quintus Metellus, consul in 109, was entrusted with the war in Africa. He waged it with obvious competence but failed to finish it and thus gave Gaius Marius, a senior officer, his chance.

Struggles of the aristocratic families

The career of Gaius Marius. Marius, born of an equestrian family at Arpinum, had attracted the attention of Scipio Aemilianus as a young soldier and, by shrewd political opportunism, had risen to the praetorship and married into the patrician family of the Julii Caesares. Though Marius had deeply offended the Metelli, once his patrons, his considerable military talents had induced Quintus Metellus to take him to Africa as a *legatus.* Marius intrigued against his commander in order to gain a consulship; he was elected (chiefly with the help of the equites and antiaristocratic tribunes) for 107 and was given charge of the war by special vote of the people. He did little better than Metellus had, but in 105 his quaestor Lucius Sulla, in delicate and dangerous negotiations, brought about the capture of Jugurtha, opportunely winning the war for Marius and Rome.

During the preceding decade a serious threat to Italy had developed in the north. Starting in 125, several Roman commanders (Marcus Flaccus has been noted) had fought against Ligurian and Gallic tribes in southern France and had finally established a Roman sphere of influence there: a road had been built linking Italy with Spain, and some garrison posts probably secured it; finally, a colony was settled at Narbonne, an important road junction (*c.* 118). But, unwilling to extend administrative responsibilities, the Senate had refused to establish a regular *provincia.* Then some migrating German tribes, chief of them the Cimbri, after defeating a Roman consul, invaded southern France, attracting native sympathy and finding little effective Roman opposition. Two more consular armies suffered defeat, and in October 105 a consul and proconsul with their forces were destroyed at Orange. There was panic in Rome, allayed only by the firm action of the other consul, Publius Rutilius Rufus.

Roman sphere of influence in southern France

At this moment news of Marius' success in Africa arrived, and he was at once dispensed from legal restrictions and again elected consul for 104. After a brilliant triumph that restored Roman morale, he took over the army prepared and trained by Rutilius. He was reelected consul year after year, while the German tribes delayed attacking Italy. Finally, in 102–101, he annihilated them at Aquae Sextiae (Aix-les-Bains) and, with his colleague, Quintus Catulus, on the Campi Raudii (near the Po delta). Another triumph and a sixth consulship (in 101) were his reward.

In his first consulship, Marius had taken a step of great (and probably unrecognized) importance: aware of the difficulties long endemic in the traditional system of recruitment, he had ignored property qualifications in enrolling his army and, as a result, had recruited ample volunteers among men who had nothing to lose. This radical solution was thenceforth generally imitated, and conscription became confined to emergencies (such as the Social and Civil wars). He also enhanced the importance of the legionary eagle (the standard), thus beginning the process that led to each legion's having a continuing corporate identity. At the same time, Rutilius introduced arms drill and reformed the selection of senior officers. Various tactical reforms in due course led to the increasing prominence of the cohort (one-tenth of a legion) as a tactical unit and the total reliance on non-Roman auxiliaries for light-armed and cavalry service. The precise development of these reforms cannot be traced, but they culminated in the much more effective armies of Pompey and Caesar.

Marius' African army had been unwilling to engage in another war, and Marius preferred to use newly levied soldiers (no longer difficult to find). But neither he nor the Senate seemed aware of any responsibilities to the veterans. In 103 a tribune, Lucius Saturninus, offered to pass a law providing land in Africa for them in return for Marius' support for some anti-oligarchic activities of his own. Marius agreed, and the large lots distributed to his veterans (both Roman and Italian) turned out to be the beginning of the Romanization of Africa. In 100, with the German wars ended, Saturninus again proved a welcome ally, arranging for the settlement of Marius' veterans in Gaul. An incidental effect was the departure of Marius' old commander and subsequent enemy, Quintus Metellus, who refused to recognize the validity of Saturninus' law and, choosing martyrdom, went into exile. But this time Saturninus exacted a high price. With his ally, the praetor Gaius Glaucia, he introduced laws to gain the favour of plebs and equites and proceeded to provide for the settlement of veterans of wars in Macedonia and Sicily in the same way as for those of Marius' war. He planned to seek reelection for 99, with Glaucia illegally gaining the consulship. Violence and even murder were freely used to accomplish these aims.

Marius now had to make a choice. Saturninus and Glaucia might secure him the continuing favour of the plebs and perhaps the equites, though they might also steal it for themselves. But as the saviour of his country and six times consul, he now hoped to become an elder statesman (*princeps*), accepted and honoured by those who had once looked down on him as an upstart. To this end he had long laboured, dealing out favours to aristocrats who might make useful allies. This was the reward Marius desired for his achievement; he never thought of revolution or tyranny. Hence, when called on to save the state from his revolutionary allies, he could not refuse. He imprisoned them and their armed adherents and did not prevent their being lynched. Yet, having saved the oligarchy from revolution, he received little reward; he lost the favour of the plebs, while the oligarchs, in view of both his birth and his earlier unscrupulous ambition, refused to accept him as their equal. Metellus was recalled; this was a bitter blow to Marius' prestige, and he preferred to leave Rome and visit Asia.

Marius' opposition to Saturninus and Glaucia

Before long a face-saving compromise was found, and Marius returned; but in the 90s he played no major part. Though he held his own when his friends and clients were attacked in the courts, his old aristocratic protégés now found more promising allies. Sulla is typical: closely associated with Marius in his early career, he was by 91 ready to take the lead in attacking Marius and (significantly) found eager support. The oligarchy could not forgive Marius.

Wars and dictatorship (c. 91–80 BC). *Events in Asia.* In foreign affairs the 90s were dominated by Asia, Rome's chief source of income. Mithradates VI, king of Pontus, had built a large empire around the Black Sea and was probing and intriguing in the Roman sphere of influence. Marius had met him and had given him a firm warning, temporarily effective: Mithradates had proper respect for Roman power. Scheming to annex Cappadocia, he had been thwarted by the Senate's instructing Sulla, as proconsul, to install a pro-Roman king there in 96–95. (It was on this occasion that Sulla received a Parthian embassy—the first contact between the two powers.) But dissatisfaction in the Roman province of Asia gave new hope to Mithradates. Ineffectively organized after annexation and corrupt in its cities' internal administration, it was soon overrun with Italian businessmen and Roman tax collectors. When the Senate realized the danger, it sent its most distinguished jurist, Quintus Mucius Scaevola (consul in 95 and *pontifex maximus*), on an unprecedented mission to reorganize Asia (94). He took Publius Rutilius Rufus—jurist, stoic philosopher, and former consul—with him as his senior officer, and after Scaevola's return Rutilius remained behind, firmly applying the new principles they had established. This caused an outcry from businessmen, whose profits Scaevola had kept within bounds; he was prosecuted for "extortion" in 92 and convicted after a trial in which Roman *publicani* and businessmen unscrupulously used their power among the class that provided criminal juries. The verdict revealed the breakdown of Gaius Gracchus' system: the class he had raised to watch over the Senate now held irresponsible power, making orderly administration impossible and endangering the empire. Various leading senators were at once vexatiously prosecuted, and political chaos threatened.

Developments in Italy. The 90s also saw dangerous developments in Italy. In the 2nd century BC, Italians as a whole had shown little desire for Roman citizenship and had been remarkably submissive under exploitation and ill-treatment. The most active of their governing class flourished in overseas business, and the more traditionally minded were content to have their oligarchic rule supported by Rome. Their admission to citizenship had been proposed as a by-product of the Gracchan reforms. By 122 it had become clear that the Roman people agreed with the oligarchy in rejecting it. The sacrifices demanded of Italy in the Numidian and German wars probably increased dissatisfaction among Italians with their patently inferior status. Marius gave citizenship to some as a reward for military distinction—illegally, but his standing (*auctoritas*) sufficed to defend his actions. Saturninus admitted Italians to veteran settlements and tried to gain citizenship for some by full admission to Roman colonies. The censors of 97–96, aristocrats connected with Marius, shared his ideas and freely placed eminent Italians on the citizen registers. This might have allayed dissatisfaction, but the consuls of 95 passed a law purging the rolls and providing penalties for those guilty of fraudulent arrogation. The result was insecurity and danger for many leading Italians. By 92 there was talk of violence and conspiracy among desperate men.

Insecurity and danger for leading Italians

It was in these circumstances that the eminent young noble, Marcus Livius Drusus, became tribune for 91 and hoped to solve the menacing accumulation of problems by means of a major scheme of reforms. He attracted the support of the poor by agrarian and colonial legislation and tried to have all Italians admitted to citizenship and to solve the jury problem by a compromise: the courts would be transferred to the Senate, and 300 equites would be admitted to it. (To cope with the increase in business it would need this expansion in size.) Some leading senators, frightened at the dangerous situation that had developed, gave weighty support. Had Drusus succeeded, the poor and the Italians might have been satisfied; the equites, deprived of their most ambitious element by promotion, might have acquiesced; and the Senate, always governed by the prestige of the noble *principes* rather than by votes and divisions, could have returned, little changed by the infusion of new blood, to its leading position in the process of government. But Drusus failed. Some members of each class affected were more conscious of the loss than of the gain; and an active consul, Lucius Philippus, provided leadership for their disparate opposition. After much violence, Drusus' laws were declared invalid. Finally he himself was assassinated. The Italians now rose in revolt (the Social War), and in Rome a special tribunal, manned by the Gracchan jury class, convicted many of Drusus' supporters until the Senate succeeded in suspending its sittings because of the military danger.

The Social War

The first year of the Social War (90) was dangerous: the tribes of central and southern Italy, traditionally among the best soldiers in Rome's wars, organized in a confederacy for the struggle that had been forced upon them. Fortunately all but one of the Latin cities—related to Rome by blood and tradition and specially favoured by Roman law—remained loyal: their governing class had for some time had the privilege of automatically acquiring Roman citizenship by holding local office. Moreover, Rome now showed its old ability to act quickly and wisely in emergencies: the consul Lucius Caesar passed a law giving citizenship to all Italians who wanted it. The measure came in time to head off major revolts in Umbria and Etruria, which accepted at once.

Civil war and the rule of Lucius Sulla. In 89 the war in central Italy was won, and Gnaeus Pompeius Strabo celebrated a triumph. Attention now turned to the East, where Mithradates had taken advantage of Rome's troubles to expel the kings of Cappadocia and Bithynia. A Roman embassy restored them, and he withdrew. However, when the envoys incited Bithynian incursions into his territory, Mithradates launched a major offensive; he overran the two kingdoms and invaded Roman territory, where he attracted the sympathy of the natives by executing thousands of Italians and defeating and capturing the Roman commanders in the area.

In Rome, various men, including Marius, had hoped for the Eastern command. But it went to Sulla, elected consul for 88 after distinguished service in the Social War. Publius Sulpicius, a tribune in that year and an old friend of Drusus, tried to continue the latter's policy of justice to the Italians by abolishing the gerrymandering that in practice deprived the new citizens of an effective vote. Finding the oligarchy firmly opposed, he gained the support of Marius (who still commanded much loyalty) for his plans by having the Eastern command transferred to him. After much street-fighting, the consuls escaped from Rome, and Sulpicius' bills were passed. Sulla's response was totally unforeseen: he appealed to the army he had led in the Social War, which was still engaged in mopping-up operations in Campania, and persuaded them to march on Rome. He occupied the city and executed Sulpicius; Marius and others escaped. Significantly, Sulla's officers left him. It was the first time a private army of citizens had occupied Rome—an effect of Marius' army reform, which had ended by creating a "client army" loyal chiefly to its commander, and of the Social War, which had made the use of force within Italy seem commonplace. The end of the republic was foreshadowed.

Sulla's occupation of Rome

Having cowed Rome into acquiescence and having passed some legislation, Sulla left for the East. Cinna, one of the consuls of 87, at once called for the overthrow of Sulla's measures. Resisted by his colleague Octavius, he left Rome to collect an army and, with the help of Marius, occupied the city after a siege. Several leading men were killed or condemned to death, Sulla and his supporters were outlawed, and (after Marius' death early in 86) another commander was sent to Asia. The policy now changed to one of reconciliation: the Social War was wound up, and the government gained wide acceptance until Cinna was killed by mutinous soldiers (84).

Sulla meanwhile easily defeated Mithradates' forces in two battles in Boeotia, took Athens, which under a revolutionary regime had declared for Mithradates, and cleared the king's army out of Greece. While negotiating with Cinna's government, Sulla also entered upon negotiations with Mithradates and, when he heard of Cinna's death, quickly made peace and an alliance with Mithradates, driving the government's commander in Asia to suicide. After wintering his troops in the rich cities of Asia, Sulla

crossed into Greece and then into Italy, where his veteran army broke all resistance and occupied Rome (82). Sulla was elected dictator and, while Italy and all the provinces except Spain were quickly reduced, began a reign of terror (the "proscriptions"), in which hundreds of his enemies or those of his adherents were killed without trial, while their property went to enrich him and his friends. Wherever in Italy he had met resistance, land was expropriated and given to his soldiers for settlement.

While the terror prevailed, Sulla used his powers to put through a comprehensive program of reform (81). Although he had twice taken Rome with a private proletarian army, he had earlier had connections with the inner circles of the oligarchy, and after Cinna's death some eminent men who had refused to collaborate with Cinna joined Sulla. By the time Sulla's success seemed certain, even most of those who had collaborated were on his side, and he was acclaimed as the defender of the nobility who had defeated an illegal revolutionary regime. His reforms aimed chiefly at stabilizing Senate authority by removing alternative centres of power. The tribunate was emasculated; the censors' powers were reduced; provincial governors were subjected to stricter Senate control; and the equites, who had been purged of Sulla's opponents by the proscriptions, were deprived of some symbols of dignity and made leaderless by the inclusion of 300 of Sulla's chief supporters in the Senate. The jury reform of Gaius Gracchus, seen by some leading senators as the prime cause of political disintegration, could now be undone, and the criminal courts could once more become a monopoly of senators.

Sulla's attempt to restore stable government

Sulla's measures were by no means merely reactionary. His program was basically that of Marcus Drusus. His overriding aim was the restoration of stable government, and this could only come from the Senate, directed by the *principes* (former consuls and those they chose to consult). Sulla accepted and even extended recent developments where they seemed useful: the Italians retained full citizenship; the system of standing criminal courts was expanded; the practice of praetors normally spending their year of office in Rome and then going to provinces for a second year was extended to consuls and became an integral part of his system. To prevent long command of armies (which might lead to careers like his own), Sulla increased the number of praetors so that, in principle and in normal circumstances, each province might have a new governor every year. As for the overriding problem of poverty, his contribution to solving it was to settle tens of thousands of his veterans on land confiscated from enemies in Italy; having become landowners, the veterans would be ready to defend the social order, in which they now had a stake, against the dispossessed.

At the beginning of 80 Sulla laid down his dictatorship and became merely consul, with the senior Metellus (Quintus Metellus Pius), a relative of his wife, as his colleague. The state of emergency was officially ended. At the end of the year, after seeing to the election of two reliable consuls, Sulla retired to Campania as a private citizen; he hoped that the restored oligarchy would learn to govern the state he had handed over to them. For 78 Marcus Lepidus, an ambitious patrician whom Sulla disliked and distrusted, was elected consul. Sulla did not intervene. Within a few months, Sulla was dead. Lepidus at once attacked his system, using the grievances of the expropriated as a rallying cry and his province of Gaul as a base. But he was easily defeated by his former colleague Quintus Catulus, assisted by young Gnaeus Pompeius (Pompey).

The Roman state in the two decades after Sulla (79–60 BC). *The early career of Pompey.* Pompey was the son of Gnaeus Pompeius Strabo, who had triumphed after the Social War but had incurred general hatred because of cold-blooded duplicity during the troubles of 88 and 87. After Strabo's death, young Pompey, who had served under him and inherited his dubiously won wealth, was protected by Cinna's government against his father's enemies. Following in his father's footsteps, he deserted the government after Cinna's death, raised a force among his father's veterans in central Italy, and helped to conquer Italy and, in a lightning campaign, Sicily and the province of Africa for Sulla. Though not old enough to hold any regular magistracy (he was born in 106), he had, from these military bases, blackmailed Sulla into granting him a triumph (81) and had married into the core of the Sullan oligarchy. Out of pique against Sulla, he had supported Lepidus' election for 78, but he had too great a stake in the Sullan system to permit Lepidus to overthrow it.

Revolt of Sertorius

Meanwhile a more serious challenge to the system had arisen in Iberia. Quintus Sertorius, a former praetor of tough Sabine gentry stock, had refused to follow most of his social betters in joining Sulla; instead he had left for Spain, where he claimed to represent the legitimate government. Although acting throughout as a Roman proconsul, with a "counter-Senate" of eminent Roman citizens, Sertorius won the enthusiastic support of the native population by his fairness, honesty, and charisma, and he soon held most of the Iberian Peninsula, defending it successfully even against a large force under Quintus Metellus Pius. When the consuls of 77 would have nothing to do with this war, Pompey was entrusted by the Senate, through the efforts of his eminent friends and sponsors, with the task of assisting Metellus. The war dragged on for years, with little glory for the Roman commanders. Although Sertorius had many sympathizers in Italy, superior numbers and resources finally wore him down, and he was assassinated by a Roman officer. Pompey easily defeated the remnants of Sertorius' forces in 72.

The death of Nicomedes IV of Bithynia (74) led to another major war. Like Attalus of Pergamum, Nicomedes left his kingdom to Rome, and this provoked Mithradates, who was in contact with Sertorius and knew of Rome's difficulties, to challenge Rome again. The Eastern command again led to intrigues in Rome. The command finally went to Lucius Lucullus, a relative of Sulla and consul in 74, who hoped to build up a countervailing power in the East.

At the same time, Marcus Antonius, father of the later Triumvir, was given a command against the pirates in the eastern Mediterranean (whom his father had already fought in 102–100), partly, perhaps, as further reinsurance against Pompey. With Italian manpower heavily committed, a minor slave rising led by Spartacus (73) assumed threatening dimensions, until Marcus Crassus (an old Sullan and profiteer in the proscriptions) volunteered to accept a special command and defeated the slaves. At this point (71) Pompey returned from Spain with his army, crucified the remnants of the slave army, and claimed credit for the victory.

Pompey and Crassus. He and Crassus now confronted each other, each demanding the consulship for 70, though Pompey had held no regular magistracy and was not a senator. Agreeing to join forces, both secured it.

During their consulship, the political, though not the administrative, part of the Sullan settlement was repealed. The tribunes' powers were fully restored; criminal juries were divided between senators and wealthy nonsenators; and, for the first time since Sulla, two censors—both supporters of Pompey—were elected, who purged the Senate and, in compiling the registers, at last fully implemented the Italians' citizenship. The year 70 also saw the prosecution of Verres (son of a "new man" and Sullan profiteer), who had surpassed the liberal Roman conventions in exploiting his province of Sicily. For future impunity he relied on his aristocratic connections (especially the Metelli and their friends), his fortune, and the known corruptibility of the Sullan senatorial juries. But Verres was unlucky. First, he had ill-treated some of Pompey's important Sicilian clients, thus incurring Pompey's displeasure; next, his case coincided with the anti-Sullan reaction of 70; finally, the Sicilians succeeded in persuading Cicero—an ambitious young "new man" from Arpinum hoping to imitate the success of his fellow citizen Marius by means of his rhetorical ability—to undertake the prosecution. Despite obstruction from Verres' friends, Cicero collected massive evidence against him, presented his case to fit into the political context of the year, and obtained Verres' conviction as an act of expiation for the shortcomings of the Sullan order.

The year 70 thus marked the loss of control by the Sul-

Loss of control by the Sullan establishment

lan establishment. The nobility (families descended from consuls) continued to gain most of the consulships, with the old patriciate (revived by Sulla after a long decline) stronger than for generations; the Senate still supervised administration and made ordinary political decisions; the system continued to rely essentially on *mos majorum* (constitutional custom) and *auctoritas* (prestige)—potent forces in the status society of the Roman Republic. The solid bases of law and power that Sulla had tried to give it had been surrendered, however. The demagogue—tribune or consul—could use the legal machinery of the popular assembly (hence such men are called *populares*), while the commander could rely on his army in the pursuit of private ambition. The situation that Sulla had tried to remedy now recurred, made worse by his intervention. His massacres and proscriptions had weeded out the defenders of lawful government, and his rewards had gone to the timeservers and the unscrupulous. The large infusion of equites into the Senate had intensified the effect. While eliminating the serious friction between the two classes, which had made the state ungovernable by 91, it had filled the Senate with men whose tradition was the opposite of that sense of mission and public service that had animated the best of the aristocracy. Few men in the new ruling class saw beyond self-interest and self-indulgence.

One result was that massive bribery and civil disorder in the service of ambition became endemic. Laws were repeatedly passed to stop them, but they remained ineffective because few found it in their interest to enforce them. Exploitation of the provinces did not decrease after Verres: governors (still with unlimited powers) feathered their own nests and were expected to provide for all their friends. Extortion cases became a political ritual, with convictions impossible to obtain. Cicero, thenceforth usually counsel for the defense, presented hair-raising behaviour as commonplace and claimed it as acceptable. The Senate's traditional opposition to annexation faded out. Pompey made Syria into a province and added a large part of Pontus to Bithynia (inherited in 74 and occupied in 70); the demagogue Clodius annexed Cyprus—driving its king to suicide—to pay for his massive grain distributions in Rome; Caesar, finally, conquered Gaul by open aggression and genocide and bled it white for the benefit of his friends and his ambitions. Crassus would have done the same with Parthia, had he succeeded. Opposition to all this in the Senate, where it appeared, was based on personal or political antagonism. If the robber barons were attacked on moral grounds, it was because of the use they made of their power in Rome.

Politically, the 60s lay under the shadow of Pompey. Refusing to take an ordinary province in 69, he waited for his chance. It came in 67 when his adherent Gabinius, as tribune, secured him, against the opposition of all important men, an extraordinary command with unprecedented powers to deal with the pirates. Pompey succeeded within a few months where Antonius and others had failed. The equites and the people were delighted because trade, including Rome's food imports, would now be secure. Meanwhile Lucullus had driven Mithradates out of Anatolia and into Armenia; but he had offended Roman businessmen by strict control and his own soldiers and officers by strict discipline. Faced with mutinies, he suffered a reverse and became vulnerable to attacks in Rome. In 66 another tribunician law appointed Pompey, fresh from his naval victories, to take over supreme command in the East, which he did at once, studiously insulting his predecessor. He quickly defeated Mithradates and procured his death, then spent some time in a total reorganization of the East, where Asia (the chief source of revenue) was protected by three further provinces and a ring of client states beyond the frontier. The whole of the East now stood in his *clientela* (clientship), and most of it owed him money as well. He returned by far the wealthiest man in Rome.

Defeat and death of Mithradates VI

Political suspicion and violence. Meanwhile Roman politics had been full of suspicion and violence, much of it stirred up by Crassus who, remembering 71, feared Pompey's return and tried to make his own power impregnable. There was much material for revolution, with poverty (especially in the country, among families dispossessed by Sulla) and debt (among both the poor and the dissolute rich) providing suitable issues for unscrupulous *populares*. One such man, the patrician Catiline, after twice failing to gain the consulship by traditional bribery and intrigue, put himself at the head of a movement planning a coup d'état in Rome to coincide with an armed rising in Italy (late 63). Cicero, as consul, defeated these efforts and, relying on the doubtful legality of a Senate vote in support, had Catiline's eminent Roman associates executed. Catiline himself fell in a desperate battle.

Isolation of Cicero

For Cicero—the "new man" who had made his way to the top by his own oratorical and political skill, obliging everyone by unstinting service, representing Pompey's interests in Rome while avoiding offense to Pompey's enemies—this was the climax of his life. Like his compatriot Marius, he had saved the state for its rulers: he had taken resolute action when those rulers were weak and vacillating; and, like Marius, he got small thanks for it. Pompey was miffed at having to share his fame with a municipal upstart, and eminent gentlemen could not forgive that upstart for having driven patricians to their death.

Pompey's return was peaceful. Like Marius, he wanted recognition, not tyranny. He dismissed his army, to the surprise of Crassus and others, and basked in the glory of his triumph and the honours voted to him. But having given up power, he found himself caught in a net of constitutional obstruction woven by his politically experienced enemies and was unable to have either of his principal demands met: land for his veterans and the ratification of his arrangements in the East. It was at this point that Caesar returned from Spain.

Gaius Julius Caesar, descended (as he insisted) from kings and gods, had shown talent and ambition in his youth: he opposed Sulla but without inviting punishment, married into the oligarchy but advocated popular causes, vocally defended Pompey's interests while aiding Crassus in his intrigues and borrowing a fortune from him, flirted with Catiline but refused to dabble in revolution, then worked to save those whom Cicero executed. In 63 he won a startling success: defeating two distinguished *principes*, he, who had not yet been praetor, was elected *pontifex maximus*—a post of supreme dignity, power, and patronage. Despite some cynicism among Roman aristocrats toward the state religion, its ceremonial was kept up and was a recognized means of political manipulation; thus priesthoods could give more lasting power than magistracies, in addition to the cachet of social success. Young Caesar was now head of the hierarchy. After his praetorship (62), Caesar successfully governed Spain, clearing a surplus sufficient to pay off his debts. On returning to Rome, he naturally hoped for the consulship of 59; but his enemies, by legal chicanery, forced him to choose between standing for office and celebrating a triumph. He gave up the triumph and easily became consul.

Alliance of Caesar, Pompey, and Crassus

The final collapse of the Roman Republic (59–44 BC). *Caesar, Pompey, and Crassus.* For his consulship Caesar fashioned an improbable alliance: his skill in having won the trust of both Crassus and Pompey enabled him to unite these two enemies in his support. Crassus had the connections, Pompey had the soldiers' vote, and Caesar was consul and *pontifex maximus*. The combination (often misleadingly called the "first Triumvirate") was invincible, especially since the consul Caesar had no scruples about countering legal obstruction with open force. Pompey got what he wanted, and so did Crassus (whose immediate need was a concession to the Asian tax farmers, in whose companies he probably had much of his capital). In return, Caesar got a special command in Cisalpine Gaul and Illyricum for five years by vote of the people; the Senate itself, on Pompey's motion, extended it to Transalpine Gaul. Marriage alliances sealed the compact, chief of them Pompey's marriage to Caesar's daughter Julia.

Caesar left for Gaul, but Rome was never the same; the shadow of the alliance hung over it, making the old-style politics impossible. In 58 Publius Clodius, another aristocratic demagogue, was tribune and defended Caesar's interests. Cicero had incurred Clodius' enmity and was now sacrificed to him: he was driven into exile as having unlawfully executed citizens in 63. By 57 Caesar's allies

had drifted back into rivalry. Pompey secured Cicero's return, and Cicero at once tried to break up the alliance by attracting Pompey to the Senate's side. Just when he seemed about to succeed, the three dynasts secretly met and revived their compact (56). Rome had to bow once more. In 55 Pompey and Crassus were consuls, and the contents of their secret agreement were slowly revealed. Caesar, whom his enemies had made efforts to recall, was prolonged in his command for five years and (it later appeared) had been promised another consulship straight after, to secure him against prosecution and give him a chance of another army command. Pompey was given a special command over all of Spain, which he exercised through deputies while he himself remained just outside Rome to keep an eye on the city. Crassus, who now needed glory and new wealth to equal those of his allies, was to attack Parthia with a large army. Thus the three dynasts would practically monopolize military power for the foreseeable future.

Cicero, among others, had to submit and was thenceforth their loyal spokesman. After his achievement of 63 he had dreamed of leading a coalition of all "right-thinking" men in Italy in defending the traditional oligarchy, but he had found little support among the oligarchy. He now used this fact to rationalize his surrender. His brother took service in Gaul under Caesar.

The dynasts' pact did not even bring peace. Clodius, as tribune, had created a private army, and there was no state force to counter it. Pompey could have done it by calling his soldiers in, but the Senate did not trust him enough to request this, and Pompey did not wish to parade himself as an unashamed tyrant. Other men formed private armies in opposition to Clodius, and one Milo at last managed to have him killed after a scuffle (52). By then, however, Roman politics had radically and unexpectedly changed.

End of the pact

Political maneuvers. Julia died in 54, breaking the ties between Caesar and Pompey. Caesar pressed Pompey to renew them, but Pompey held off, preserving his freedom of action. Crassus' Parthian campaign ended in disaster and in Crassus' death (53). By 52 Pompey and Caesar stood face to face, still nominally friends but with no personal link between them and no common interests. Caesar, by conquering the whole of Gaul, had almost equaled Pompey's prestige and, by his utterly ruthless way of waging war, Pompey's wealth. Unlike Pompey, he used his wealth to dispense patronage and buy useful friends. At this point Pompey cautiously offered the oligarchy his support. It had much to give him that he wanted—control of the administrative machine, respectability, and the seal of public approval. Its leaders (even the intransigent young Cato, who had led opposition to the three individually long before their alliance and to their joint oppression of the state ever since) now recognized that acceptance of Pompey's terms and surrender to his protection was their only chance of survival. Pompey at once turned firmly against Milo, who presented a political threat: if Milo could use the force that had killed Clodius to keep firm control of Rome, he—an ambitious man of known conservative views—might in due course offer an alternative and more trustworthy champion to the oligarchy. But he was not yet ready. Pompey forced them to make their choice at once, and they chose Pompey in preference. He was made sole consul and had Milo convicted by an intimidated court. Meanwhile he had made a marriage alliance with the noblest man in Rome, Quintus Metellus Scipio, who became his colleague in the consulship. The state had captured Pompey (or vice versa), and Caesar stood alone in opposition to both of them. During the next two years there were a series of maneuvers: the Senate leaders, with Pompey's silent support, worked for Caesar's recall, which would have meant his instantly sharing the fate of Milo; while Caesar and his agents in Rome tried to strike some bargain that would ensure his safety and his future in politics. Finally, Pompey declared himself, and, early in 49, the Senate voted to outlaw Caesar. Two tribunes supporting him (one of them Mark Antony) had to flee. By the time they reached him, Caesar had already crossed the Rubicon: he now had a cause.

Civil war. Pompey had exuded confidence over the outcome if it came to war. In fact, however, Caesar's veterans were unbeatable, and both men knew it. To the disgust of his followers, Pompey evacuated Rome, then Italy. His plan was to bottle Caesar up in Italy and starve him out. But Caesar, in a lightning sweep, seized Massilia and Spain from Pompey's commanders, then crossed into Greece, where a short campaign ended in Pompey's decisive defeat at Pharsalus (48). Pompey fled to Egypt, where he was assassinated by a man hoping thus to curry Caesar's favour. This was by no means the end of the war. Almost at once Caesar was nearly trapped at Alexandria, where he had intervened in a succession dispute; but he escaped and installed Cleopatra on the throne, for personal as well as political reasons. In Africa the Pompeian forces and their native allies were not defeated until Caesar himself moved against them and annihilated them at Thapsus. Cato, disdaining the victor's pardon, committed suicide at Utica (46). In Spain, where Pompey's name was still powerful, his sons organized a major rising, which Caesar himself again had to defeat at Munda (45) in the bloodiest battle of the war. By the time he returned, he had only a few months to live.

The dictatorship and assassination of Caesar. In Rome the administrative machine had inevitably been disrupted, and Caesar had always remained in control, as consul or as dictator. Those who had feared proscriptions, or hoped for them, were proved wrong. Some of Caesar's enemies had their property confiscated, but it was sold at fair value; most were pardoned and suffered no loss. One of these was Cicero, who, after much soul-searching, had followed his conscience by joining Pompey before Pharsalus.

Changes under Caesar

Poverty and indebtedness were alleviated, but there was no wholesale cancellation of debts or redistribution of property, and many of Caesar's adherents were disappointed. Nor was there a general reform of the republic. (Caesar's only major reform was of the calendar: indeed, the Julian calendar proved adequate for centuries.) The number of senators and magistrates was increased, the citizenship was more freely given, and the province of Asia was relieved of some of its tax burden. But Caesar had no plan for reforming the system—not even to the extent that Sulla had tried to do, for Sulla had at least planned for his own retirement. For a time, honourable men, such as Cicero, hoped that the "Dictator for Settling the Constitution" (as Caesar called himself) would produce a real constitution—some return to free institutions. By late 45 that hope was dead. Caesar was everywhere, doing everything to an almost superhuman degree. He had no solution for the crisis of the republic except to embody it in himself and none at all for the hatred of his peers, which he knew this was causing. He began to accept more and more of the honours that a subservient Senate invidiously offered, until finally he reached a position perilously close to kingship (an accursed term in Rome) and even deification. Whether he passed those hazy boundary lines is much debated and not very important. He had put himself in a position in which no Roman ought to have been and which no Roman aristocrat could tolerate. As a loyal friend of his was later to say: "With all his genius, he saw no way out." To escape the problem or postpone it, he prepared for a Parthian war to avenge Crassus—a project most likely to have ended in similar disaster. Before he could start on it, about 60 men—former friends and old enemies, honourable patriots and men with grievances—struck him down in the Senate on March 15, 44 BC.

The Triumvirate and Octavian's achievement of sole power. Brutus and Cassius, the organizers of the conspiracy, expected all Romans to rejoice with them in the rebirth of "freedom." But to the Roman people the freedom of the governing class had never meant very much; the armies (especially in the west) were attached to Caesar; and the Senate was full of Caesarians at all levels, cowed but biding their time. Mark Antony, the surviving consul, whom Brutus had been too scrupulous to assassinate with his master, gradually gained control of the city and the official machinery, and the "liberators" withdrew to the East. But a challenger for the position of leader of the Caesarians soon appeared in the person of Octavian, Caesar's son by adoption and now his heir. Though not

yet 20, Octavian proved an accomplished politician; he attracted loyalty as a Caesarian while cooperating against Antony with the Senate, which, under Cicero's vigorous leadership, now turned against the consul. Cicero hoped to fragment and thus defeat the Caesarian party, with the help of Brutus and Cassius, who were making good progress in seizing control of the eastern provinces and armies. In 43 the two consuls (both old Caesarian officers) and Octavian defeated Antony at Mutina, and success seemed imminent. But the consuls died, and Octavian demanded and, by armed force, obtained the consulship; and the armies of Italy, Spain, and Gaul soon showed that they would not fight against one another. Octavian, Antony, and Lepidus (the senior Caesarian with an army) now had themselves appointed "Triumvirs for Settling the Constitution" for five years and secured control of Italy by massive proscriptions and confiscations (Cicero, Antony's chief enemy, was among the first to die). They then defeated and killed Brutus and Cassius at Philippi (42) and divided the Roman world among themselves, with Lepidus, a weak man accidentally thrust into prominence, getting the smallest share. Octavian, who was to control Italy, met armed opposition from Antony's brother and wife, but they got no help from Antony and were defeated at Perusia (41). Octavian and Antony sealed their alliance with a marriage compact: Antony married Octavia, Octavian's sister. Octavian then confronted Pompey's son Sextus Pompeius, who had seized control of the islands off Italy. After much diplomatic maneuvering (including another meeting with Antony), Octavian attacked and defeated Sextus; when Lepidus tried to reassert himself, Octavian crushed him and stripped him of his office of Triumvir (while with conspicuous piety leaving him the chief pontificate, now an office without power). Octavian now controlled the West and Antony the East, still officially as Triumvirs (their term of office had been extended), even though Lepidus had been eliminated in 36.

Formation of the Triumvirate

Each of the two leaders embarked on campaigns and reorganization in his half—Octavian in Illyricum, Antony particularly on the Parthian frontier. But Antony now married Cleopatra and tried to make Egypt his military and political base. In a war of propaganda, Octavian gradually convinced the western provinces, Italy, and most of the Roman upper class that Antony was sacrificing Roman interests, trying to become a Hellenistic king in Alexandria, and planning to rule the Roman world from there with Cleopatra. In 32, though he now held no legal position, Octavian intimidated most of Antony's remaining aristocratic friends into joining him, made the whole West swear allegiance to himself, and in 31, as consul, crossed into Greece to attack Antony. On September 2 he defeated Antony and Cleopatra in a naval battle at Actium. Though in itself not a major victory, it was followed by the disintegration of Antony's forces, and Antony and Cleopatra finally committed suicide in Alexandria (30).

(E.Ba./R.P.Sa.)

INTELLECTUAL LIFE OF THE LATE REPUBLIC

The late Roman Republic, despite its turmoil, was a period of remarkable intellectual ferment. Many of the leading political figures were men of serious intellectual interests and literary achievement; foremost among them were Cicero, Caesar, Cato, Pompey, and Varro, all of them senators. The political upheaval itself leavened intellectual life; imperial senators were to look back to the late republic as a time when great political struggles stimulated great oratory, something the more ordered world of the emperors could no longer do.

The seeds of intellectual development had been sown in the late 3rd and early 2nd centuries; the flowering came in the last generation of the republic. As late as the 90s BC the Romans still appear relatively unsophisticated. Greek intellectuals were absorbed in debates among themselves, giving only passing nods to Romans by dedicating untechnical works to them. In 92 the censors issued an edict closing down the schools of Latin rhetoric in Rome. Serious students such as Cicero had to go east in the 80s to receive their higher education from leading Greek philosophers and rhetoricians.

The centre of intellectual life began to shift toward the West after the 90s. As a result of the Mithradatic wars, libraries were brought from the East to Italy. The Hellenistic kingdoms, which had provided the patronage for much intellectual activity, were dismantled by Pompey and Octavian, and Greek intellectuals increasingly joined the retinues of great Roman senators such as Pompey. Private Roman houses, especially senatorial villas on the Bay of Naples, became the focus of intellectual life; it was there that libraries were reassembled and Greek teachers kept as dependents.

Roman disciplines

Roman traditions favoured the development of certain disciplines, creating a pattern that was distinct from the Greek. Disciplines related to the public life of senators prospered—notably oratory, law, and history; certain fields of study were judged fit for diversions in leisure hours, and still others were considered beneath the dignity of an honourable Roman. Areas such as medicine and architecture were left to Greeks and others of lower status, and mathematics and the sciences aroused little interest. Greek slaves especially played an important role in the intellectual life of the late republic, serving in roles as diverse as teachers, copyists of manuscripts, and oral readers to aristocrats.

By the beginning of the imperial era the maturing of Roman intellectual culture was evident. Caesar had commissioned Varro to organize the first public library in Rome, and Greek scholars such as the geographer Strabo moved west to pursue their studies in Rome.

Grammar and rhetoric. The education of the Roman elite was dominated by training in language skills, grammar, and rhetoric. The *grammatici*, who taught grammar and literature, were lower-class and often servile dependents. Nevertheless, they helped to develop a Roman consciousness about "proper" spelling and usage that the elite adopted as a means of setting themselves off from humbler men. This interest in language was expressed in Varro's work on words and grammar, *De Lingua Latina* (43?), with its prescriptive tone. Rhetoric, though a discipline of higher status, was still taught mainly by Greeks in Greek. The rhetoricians offered rules for composition: how to elaborate a speech with ornamentation and, more important, how to organize a work through the dialectical skills of definition and division of the subject matter into analytical categories. The Romans absorbed these instructions so thoroughly that the last generation of the republic produced an equal of the greatest Greek orators in Cicero. The influence on Roman culture of dialectical thinking, instilled through rhetoric, can hardly be overstated; the result was an increasingly disciplined, well-organized habit of thinking. This development can be seen most clearly in the series of agricultural works by Roman authors: whereas Cato's 2nd-century *De agricultura* is rambling and disorganized, Varro's three books on *Res rusticae* (37), with their division of soils into 99 types, seem excessively organized.

Law and history. Roman law, though traditional in content, was also deeply influenced by Greek dialectic. For centuries the law had been passed down orally by pontifical priests. It emerged as an intellectual discipline only in the late republic, when men who saw themselves as legal specialists began to write treatises aimed at organizing existing law into a system, defining principles and concepts, and then applying those principles systematically. Quintus Mucius Scaevola was a pivotal figure: a pontifex in the traditional role, he published the first systematic legal treatise, *De iure civili*, in the 80s. Cicero credited his contemporary Servius Sulpicius Rufus with being the jurist who transformed law into a discipline (*ars*).

The decisive events of the late republic stimulated the writing of history. The first extant historical works in Latin (rather than in Greek) date from this period: Sallust's *Bellum Iugurtinum* (*Iugurthine War*) and *Bellum Catilinae* (*Catilinarian Conspiracy*) and Caesar's memoirs about his Gallic and civil wars. The rapid changes also prompted antiquarian studies as Roman senators looked back to archaic institutions and religious rituals of the distant past to legitimize or criticize the present. Varro's 41 books (now lost) on *Antiquitates terum humanarum et divinarum* ("Antiquities of things human and divine") were

influential in establishing the traditions of early Rome for future generations.

Philosophy and poetry. Philosophy and poetry were suitable as pastimes for senators; few, however, were as serious about philosophy as the younger Cato and Cicero. Even Cicero's philosophical works were not technical treatises by Greek standards; rather, they were presented as dialogues among leading senators in their leisure. Similarly, Lucretius' *De rerum natura* (*On the Nature of Things;* 50s) offered, in verse, a nontechnical explanation of Epicureanism. The technical philosophical works were written by humbler men and are now lost. A survey of their names and titles, however, shows that stoicism was not yet the dominant philosophical school it later became; more in evidence were the Epicureans, peripatetics, and academics. There also were revivals of Aristotelian and Pythagorean studies in this period.

The best-known poets of the late republican and civil war periods came from well-to-do Italian families. Catullus from Verona (*c.* 84–*c.* 54) had a reputation as *doctus* (learned) for his exquisitely crafted poems full of literary allusions in the Alexandrian style. Far from cumbersome, however, were many of his short, witty poems that challenged traditional Roman mores and deflated senatorial pretensions. Rome's greatest poets, Virgil (70–19) and Horace (65–8), were born during the republic, came of age during the civil wars, and survived to celebrate the victory of their patron, Augustus. Virgil's *Eclogues* one and nine, written during the civil wars, poignantly evoke the suffering of the upheaval that ironically inspired Rome's highest intellectual and artistic achievements. (R.P.Sa.)

THE EARLY ROMAN EMPIRE (31 BC–AD 193)

The consolidation of the empire under the Julio-Claudians. *The establishment of the principate under Augustus.* Actium left Octavian the master of the Roman world. This supremacy, successfully maintained until his death more than 40 years later, made him the first of the Roman emperors. Suicide removed Antony and Cleopatra and their potential menace in 30 BC, and the annexation of Egypt with its Ptolemaic treasure brought financial independence. With these reassurances Octavian could begin the task of reconstruction.

The first of the Roman emperors

Law and order had vanished from the Roman state when its ruling aristocrats refused to curb their individual ambitions, when the most corrupt and violent persons could gain protection for their crimes by promising their support to the ambitious, and when the ambitious and the violent together could thus transform a republic based on disciplined liberty into a turbulent cockpit of murderous rivalries. Good government depended on limits being set to unrestrained aspirations, and Octavian was in a position to impose them. But his military might, though sufficiently strong in 31 BC to guarantee orderly political processes, was itself incompatible with them; nor did he relish the role of military despot. The fate of Julius Caesar, an eagerness to acquire political respectability, and his own esteem for ancestral custom combined to dissuade Octavian from it. He wished to be, in his own words, "the author of the best civilian government possible." His problem was to regularize his own position so as to make it generally acceptable, without simultaneously reopening the door to violent lawlessness. His pragmatic responses not only ensured stability and continuity but also respected republican forms and traditions so far as possible.

Large-scale demobilization allayed people's fears; regular consular elections raised their hopes. In 29–28 BC Octavian carried out, with Marcus Vipsanius Agrippa, his powerful deputy, the first census of the Roman people since 70; and this involved drawing up an electoral roll for the Centuriate Assembly. Elections followed, and Octavian was inevitably chosen consul. Then, on Jan. 13, 27 BC, he offered to lay down his powers. The Roman Senate rejected this proposal, charging him instead to administer (besides Egypt) Spain, Gaul, and Syria for the next 10 years, while it itself was to supervise the rest of the empire. Three days later, among other honours, it bestowed upon him the name by which he has ever since been known, Augustus.

As most of the troops still under arms were in the regions entrusted to Augustus' charge, the arrangements of 27 BC hardly affected his military strength. Moreover, so long as he was consul (he was reelected every year until 23 BC), he was civilian head of government as well. In other words, he was still preeminent and all-powerful, even if he had, in his own words, placed the *res publica* at the disposal of the Senate and the Roman people. Augustus particularly wished to conciliate the senatorial class, without whose cooperation civilian government was impossible. But his monopolization of the consulship offended the Senate, making a different arrangement clearly necessary. Accordingly, in 23 Augustus made a change; he vacated the consulship and never held it again (except momentarily in 5 BC and again in 2 BC, for a limited, specific purpose). In its place he received the tribunician power (*tribunicia potestas*). He could not become an actual plebeian tribune, because Julius Caesar's action of making him a patrician had disqualified him for the office. But he could acquire the rights and privileges pertaining to the office; and they were conferred upon him, apparently by the Senate, whose action was then ratified by the popular assembly. He had already been enjoying some of a tribune's privileges since 36; but he now acquired them all and even some additional ones, such as the right to convene the Senate whenever he chose and to enjoy priority in bringing business before it. Through his tribunician power he could also summon the popular assembly and participate fully in its proceedings. Clearly, although no longer consul, he still retained the legal right to authority in civilian affairs.

Augustus' tribunician power

The arrangement of 23 entailed an additional advantage. The power of the plebeian tribune was traditionally associated with the protection of citizens, and Augustus' acquisition of it was therefore unlikely to rouse resentment. Indeed, Augustus thenceforth shrewdly propagated the notion that, if his position in the state was exceptional (which it clearly was), it was precisely because of his tribunician power. Although he held it for only one year at a time, it was indefinitely renewable and was pronounced his for life. Thus, it was both annual and perpetual and was a suitable vehicle for numbering the years of his supremacy. His era (and this is true also of later emperors) was counted officially from the year when he acquired the tribunician power.

The year 23 likewise clarified the legal basis for Augustus' control of his *provincia* (the region under his jurisdiction) and its armed forces. The Senate invested him with an *imperium proconsulare* (governorship and high command), and, while this had a time limit, it was automatically renewed whenever it lapsed (usually every 10 years). This proconsular *imperium,* furthermore, was pronounced valid inside Italy, even inside Rome and the *pomerium* (the boundary within which only Roman gods could be worshiped and civil magistrates rule), and it was superior (*majus*) to the *imperium* of any other proconsul. Thus, Augustus could intervene legally in any province, even in one entrusted to someone else.

The network of favours owed him that Augustus had cultivated within the state, among people of the greatest authority over their own networks, made his position virtually unassailable, but he avoided provoking this high class of his supporters, senatorial and equestrian, by not drawing attention to the most novel and autocratic of the many grants of power he had received, the *imperium proconsulare majus.* Instead, he paraded the tribunician power as the expression of his supreme position in the state.

After 23 no fundamental change in Augustus' position occurred. He felt no need to hold offices that in republican times would have conferred exceptional power (*e.g.,* dictatorship, lifetime censorship, or regular consulship), even though these were offered him. Honours, of course, came his way: in 19 BC he received some consular rights and prerogatives, presumably to ensure that his *imperium* was in no particular inferior to a consul's; in 12, when Lepidus died, he became *pontifex maximus* (he had long since been elected into all of the priestly colleges); in 8 BC the 8th month of the year was named after him; in 2 BC he was designated *pater patriae* ("father of his country"), a distinction that he particularly esteemed because it sug-

Augustus' honours

gested that he was to all Romans what a *paterfamilias* was to his own household. He also accepted special commissions from time to time: *e.g.*, the supervision of the supply of grain and water, the maintenance of public buildings (including temples), the regulation of the Tiber, the superintendence of the police and fire-fighting services, and the upkeep of Italy's roads. Such behaviour advertised his will and capacity to improve the lives of people dependent on him. Of that capacity, manifest on a grand scale, his tribunician power and proconsular *imperium* were only the formal expression. He was a charismatic leader of unrivaled prestige (*auctoritas*), whose merest suggestions were binding.

Like an ordinary Roman, he contented himself with three names. His, however, Imperator Caesar Augustus, were absolutely unique, with a magic all their own that caused all later emperors to appropriate them, at first selectively but after AD 69 in their entirety. Thereby they became titles, reserved for the emperor (or, in the case of the name Caesar, for his heir apparent); from them derive the titles emperor, kaiser, and tsar. Yet, as used by Augustus and his first four successors, the words Imperator Caesar Augustus were names, not titles—that is, respectively, praenomen, nomen (in effect), and cognomen. One title that Augustus did have was princeps (prince); this, however, was unofficial—a mere popular label, meaning Rome's first citizen—and government documents such as inscriptions or coins do not apply it to Augustus. But because of it the system of government he devised is called the principate.

The Roman Senate and the urban magistracies. Augustus regarded the Senate, whose leading member (*princeps senatus*) he had become in 28, as a body with important functions; it heard fewer overseas embassies than formerly, but otherwise its dignity and authority seemed unimpaired; its members filled the highest offices; its decrees, although not formally called laws, were just as binding; it soon became a high court, whose verdicts were unappealable; it supervised the older provinces and nominally the state finances as well, and it also in effect elected the urban magistrates; formally, even the emperor's powers derived from the Senate. Nevertheless, it lacked real power. Its provinces contained few troops (and by AD 40 it had ceased to control even these few). Hence, it could hardly dispute Augustus' wishes. In fact, real power rested with Augustus, who superintended state finances and above all controlled membership in the Senate; every senator's career depended on his goodwill. But he valued the Senate as the repository of the true Roman spirit and traditions and as the body representing public opinion. He was considerate toward it, shrewdly anticipated its reactions, and generally avoided contention with it. He regularly kept it informed about his activities; and an imperial council (Consilium Principis), which he consulted on matters of policy, in the manner of a republican magistrate seeking the opinion of his advisory committee, consisted of the consuls, certain other magistrates, and 15 senators—not handpicked by him but chosen by lot every six months.

Augustus' relationship with the Senate

To rid the Senate of unworthy members, he reduced its numbers by successive reviews to about 600 (from the triumviral 1,000 or more). Sons of senators and men of good repute and substance who had served in the army and the *vigintiviri* ("board of twenty," minor magistracy) could become members by being elected, at age 25 or over, to the quaestorship. Their subsequent rank in the Senate depended on what other magistracies they managed to win; these were, in ascending order, the aedileship (or plebeian tribunate), the praetorship, and the consulship. No one disliked by Augustus could expect to reach any of them, while anyone whom he nominated or endorsed was sure of election. Despite the emperor's control, there were usually enough candidates for keen contests. By AD 5 *destinatio* seems to have been the practice—that is, a special panel of senators and equites selected the praetors and consuls, and the *comitia centuriata* automatically ratified their choice. In about AD 5, likewise, the consulship was shortened to six months. This not only gratified senators and increased the number of high-ranking qualified officials but also showed that the consuls' duties were becoming largely ceremonial. This was also true, but to a far lesser degree, of the other unpaid magistrates. A senator really made his mark in between his magistracies, when he served in important salaried posts, military or civilian or both, sometimes far from Rome.

The equestrian order. Senators, however, were either too proud or too few to fill all the posts. Some posts were considered menial and went to the emperor's freedmen or slaves. Others were entrusted to equites, and the equestrian order soon developed into one of the great institutions of the empire. Augustus decided that membership in the order should be open to Roman citizens of means and reputation but not necessarily of good birth. Ultimately, there were thousands of equites throughout the empire. Although this was a lower aristocracy, a good career was available to them. After tours of duty as an army officer (the so-called *militiae equestres*), an aspiring eques might serve as the emperor's agent (procurator) in various capacities and eventually become one of the powerful prefects (of the fleet, of the *vigiles*, or fire brigade, of the grain supply, of Egypt, or of the Praetorian Guard). This kind of an equestrian career became standardized only under Claudius I; but Augustus began the system and, by his use of equites in responsible posts, founded the imperial civil service, which later was headed chiefly by them. The equites also performed another function: the senatorial order had difficulty in maintaining its numbers from its own ranks and depended on recruitment from below, which meant from the equestrian order. Because this order was not confined to Rome or even to Italy, the Senate gradually acquired a non-Italian element. The western provinces were already supplying senators under Augustus.

Rise of equites to a privileged status

Administration of Rome and Italy. Ordinary Roman citizens who were neither senators nor equites were of lesser consequence. Although still used, the old formula *senatus populusque Romanus* ("the Senate and the Roman people") had changed its meaning: in effect, its *populusque Romanus* portion now meant "the emperor." The "Roman people" had become the "Italian people," and it was embodied in the person of Augustus, himself the native of an Italian town. To reduce the risk of popular demonstrations in Rome, the emperor provided grain doles, occasional donatives, and various entertainments; but he allowed the populace no real power. After AD 5 the Roman people's participation in public life consisted in the formality of holding occasional assemblies to ratify decisions made elsewhere. Ultimately, this caused the distinction between the Roman citizens of Italy and the provincial inhabitants of the overseas empire to disappear; under Augustus, however, the primacy of Italy was insistently emphasized.

Indeed, Italy and justice for its inhabitants were Augustus' first cares. Arbitrary triumviral legislation was pronounced invalid after 29 BC, and ordinary Roman citizens everywhere had access to Augustus' own court of appeal (his appellate jurisdiction dated from 30 BC and in effect replaced the republican appeal to the people). His praetorian and urban cohorts provided physical security; his officials assured grain supplies; and he himself, with help from such aides as Agrippa, monumentalized Italian towns. The numerous Augustan structures in Italy and Rome (as he boasted, a city of brick before his time and of marble afterward) have mostly perished, but impressive ruins survive (*e.g.*, aqueduct, forum, and mausoleum in Rome; bridge at Narni; arch at Fano; gate at Perugia). Doubtless their construction alleviated unemployment, especially among the proletariat at Rome. But economic considerations did not influence Augustus' policies much (customs tariffs, for instance, were for fiscal, not protective, purposes), nor did he build harbour works at Ostia, Rome's port. Italian commerce and industry—notably fine pottery, the so-called *terra sigillata*, and wine—nevertheless flourished in the conditions he created. Public finances, mints, and coinage issues, chaotic before him, were placed on a sound basis, partly by the introduction of a sales tax and of a new levy (inheritance taxes) on Roman citizens—who hitherto had been subject only to harbour dues and manumission (see below) charges—and partly by means of repeated subventions to the public treasury (*aerarium Saturni*) from

The primacy of Italy

Augustus' own enormous private resources (*patrimonium Caesaris*). His many highways also contributed to Italy's economic betterment.

Augustus' great achievement in Italy, however, was to restore morale and unify the country. The violence and self-aggrandizement of the 1st century BC had bred apathy and corruption. To reawaken a sense of responsibility, especially in official and administrative circles, Augustus reaffirmed traditional Italian virtues (by laws aimed against adultery, by strengthening family ties, and by stimulating the birth rate) and revived ancestral religion (by repairing temples, building new shrines, and reactivating moribund cults and rituals). To infuse fresh blood and energy into disillusioned Roman society, he promoted the assimilation of Italy: the elite of its municipal towns entered the Roman Senate, and Italy became firmly one with Rome. To keep the citizen body pure, he made manumission of slaves difficult, and from those irregularly manumitted he withheld the citizenship.

Administration of the provinces. Sharply distinguished from Italy were the provinces of the empire. From 27 BC on they were of two types. The Senate supervised the long-established ones, the so-called public provinces: their governors were chosen by lot, usually served for a year, commanded no troops, and were called proconsuls (although only those superintending Asia and Africa were in fact former consuls, the others being former praetors). The emperor supervised all other provinces, and collectively they made up his *provincia:* he appointed their governors, and these served at his pleasure, none with the title of proconsul because in his own *provincia* proconsular *imperium* was wielded by him alone. These imperial provinces might be "unarmed," but many of them were garrisoned, some quite heavily. Those containing more than one legion were entrusted to former consuls and those with a legion or less to former praetors; in both cases their governors were called *legati Augusti pro praetore* ("legates of Augustus with authority of a praetor"). There were also some imperial provinces governed not by senators but by equites (usually styled procurators but sometimes prefects); Judaea at the time of Christ's crucifixion was such an equestrian province, Pontius Pilate being its governor. An entirely exceptional imperial province was Egypt, so jealously guarded that no senator could visit it without express permission; its prefect was unique in being an equestrian in command of legions.

Two types of provinces

The provinces paid tribute, which helped to pay for the armed services, various benefactions to supporters, a growing palace staff, and the public-works programs. Periodical censuses, carefully listing provincial resources, provided the basis for the two direct taxes: *tributum soli,* exacted from occupiers of provincial soil, and *tributum capitis,* paid on other forms of property (it was not a poll tax, except in Egypt and in certain backward areas). In addition, the provinces paid indirect taxes, such as harbour dues. In imperial provinces the direct taxes (*tributa*) were paid to the emperor's procurator, an equestrian official largely independent of the governor. In senatorial provinces, quaestors supervised the finances; but, increasingly, imperial procurators also appeared. The indirect taxes (*vectigalia*) were still collected by *publicani,* who were now much more rigorously controlled and gradually replaced by imperial civil servants.

To reward his troops after faithful service, Caesar had settled them on lands mostly in the provinces, in veteran towns; and Augustus, for the same reason and to reduce the dangerous military presence in the state generally, resorted to the same procedure on a vast scale. Thus, in the space of a single generation, more than 120 new centres were organized across the empire in an explosion of urbanizing energy never equaled or even approached in later times. In the settlements called *coloniae* all residents were to be Roman citizens, and the form of government and many other aspects of life specified in their charters bore a thoroughly Roman character. Some *coloniae,* in further approximation to Italian models, enjoyed exemption from tribute. In the *municipia* only persons elected as magistrates were awarded Roman citizenship (after Hadrian, in Africa, admission was sometimes extended to the whole of the local senate); but the whole of the local aristocracy in the course of time would be in this way gradually incorporated fully into the state. In *municipia,* too, charters specified Roman forms of government. Urban centres that were wholly noncitizen, called *civitates,* enjoyed autonomy in their own affairs, under the governor's eye; they paid taxes and administered the rural territory around them. In the west, many of them were eventually granted the status of *municipia,* and they adopted the originally Italian magistracies (*duoviri* and *aediles,* collectively *quattuorviri*) and senate (*curia* or *ordo*), normally numbering 100 members. The entire West rapidly came under the administration of urban centres of these three forms, without which the central government could never have done its job. Moreover, these centres radiated economic and cultural influence around them and so had an immense effect, particularly on the way of life of the more backward areas. In the east, however, urban centres, though equally important for government purposes, had already been in existence and long settled into their own culture and forms of government.

Types of urban centres

The provinces were generally better off under the empire. Appointment over them as governor was now and henceforth generally granted with the emperor's approval. Because he thought of himself as in some ways the patron and defender of the provincial population, lax or extortionate officials could expect some loss of imperial favour, an end to their careers, or an even more severe punishment.

Emperor worship. For this priceless gift of peace many individuals and even whole communities, in Italy and elsewhere, expressed their thanks spontaneously by worshiping Augustus and his family. Emperor worship was also encouraged officially, however, as a focus of common loyalty for the polyglot empire. In the provinces, to emphasize the superiority of Italy, the official cult was dedicated to Roma et Augustus; to celebrate it, representatives from provincial communities or groups of communities met in an assembly (Consilium Provinciae), which incidentally might air grievances as well as satisfactions. (This system began in the Greek-speaking provinces, long used to wooing their rulers with divine honours. It penetrated the west only slowly, but from 12 BC an assembly for the three imperial Gallic provinces existed at Lugdunum.) In Italy the official cult was to the *genius Augusti* (the life spirit of his family); it was coupled in Rome with the Lares Compitales (the spirits of his ancestors). Its principal custodians (*seviri Augustales*) were normally freedmen. Both the Senate and the emperor had central control over the institution. The Senate could withhold a vote of posthumous deification, and the emperor could acknowledge or refuse provincial initiatives in the establishment of emperor worship, in the construction for it, or in its liturgical details. The energy, however, that infused emperor worship was to be found almost wholly among the local nobilities.

The army. It was Augustus' soldiers, however, not his worshipers, who made him all-powerful. Their allegiance, like the name Caesar, was inherited from his "father," the deified Julius. The allegiance was to the emperor personally, through a military oath taken in his name every January 1; and the soldiers owed it after his death to his son or chosen successor. This preference of theirs for legitimacy could not be ignored because they were now a standing army, something that the republic had lacked. Demobilization reduced the 60 legions of Actium to 28, a number hardly sufficient but all that Augustus' prudence or economy would countenance. These became permanent formations, each with its own number and name; the soldiers serving in them were called legionaries. Besides the legionaries there was a somewhat smaller body of auxiliaries, or supporting troops. The two corps together numbered more than a quarter of a million men. To them must be added the garrison of Italy—the praetorian cohorts, or emperor's bodyguard, about 10,000 strong—and the marines of the imperial fleet, which had its main headquarters at Misenum and Ravenna in Italy and subsidiary stations and flotillas on seas and rivers elsewhere (the marines, however, were not reckoned good combat forces). All these troops were long-service professionals—the praetorians serving 16 years; legionaries, 20; auxil-

Allegiance of the army to the emperor

iaries, 25; and marines, 28—with differing pay scales, the praetorians' being the highest. In addition to their pay, the men received donatives, shares of booty, and retirement bonuses from a special treasury (*aerarium militare*) established in AD 6 and maintained out of the sales tax and Roman citizens' death duties. Under Augustus the praetorians were normally Italians, but many legionaries and virtually all auxiliaries were provincials, mainly from the imperial provinces in the west, the legionaries coming from municipal towns and the auxiliaries from tribal areas. The tendency to use provincials grew, and by the year 100 the Roman imperial army was overwhelmingly non-Italian. Nevertheless, it helped greatly to Romanize the empire. The legionaries were Roman citizens from the day they enlisted, if not before, and the auxiliaries (after Claudius anyway) from the day they were discharged; and, though serving soldiers could not legally marry, many had mistresses whose children often became Roman citizens. The troops, other than praetorians and marines, passed their years of service in the "armed" imperial provinces—the auxiliaries in forts near the frontier and the legionaries at some distance from it in camps that showed an increasing tendency, especially after AD 69, to become permanent (some of them, indeed, developed into great European cities). There was no central reserve, because, although desirable for emergencies, it might prove dangerous in peacetime.

The officers were naturally Roman citizens. In the legions those of the highest rank (*legati* and *tribuni*) were senators or equites; lower officers (*centuriones*) might enter directly from Italian or provincial municipalities or might rise through the ranks; by the time they retired, if not sooner, many of them were equites. In the auxiliaries the unit commanders (*praefecti*) were equites, often of provincial birth. On retirement the soldiers frequently settled in the provinces where they had served, made friends, and perhaps acquired families. Imperial policy favoured this practice. Thus the army, which had done much to introduce into the provinces Romans of all ranks, with their own way of life, through veteran settlements of the 40s, 30s, and 20s BC, continued in the same role on a more modest and casual scale throughout the Augustan reign and for two centuries or so afterward.

Foreign policy. After Actium and on two other occasions, Augustus solemnly closed the gates of the shrine of Janus (a gesture of peace) to show that Rome had peace as well as a *princeps*. These well-publicized gestures were purely temporary; the gates were swiftly reopened. His proconsular *imperium* made Augustus the arbiter of peace and war, and an ostensible search for defensible frontiers made his a very warlike reign. While the republic had left the limits of Roman territorial claims rather vague and indefinite, he planned conquests stretching to the boundaries defined by nature (deserts, rivers, and ocean shores), not always, however, with immediate annexation in mind. When annexation did occur, it was followed by the construction of solidly built military roads, paved with thick stone blocks: these also served the official post system (*cursus publicus*) and were provided with rest stages and overnight lodges at regular intervals.

Territorial expansion under Augustus

Areas where subjugation looked arduous and where Romanization seemed problematic were left to client kings, dependent on the emperor's support and goodwill and under obligation to render military aid to Rome. Such satellite kingdoms spared Augustus the trouble and expense of maintaining strong defenses everywhere; nevertheless, their ultimate and intended destiny was incorporation as soon as it suited their overlord's convenience. Usually, territory was gained more easily by creating and subsequently incorporating a client kingdom than by launching an expansionist war.

In the south, Augustus found suitable frontiers quickly. In 25 BC an expedition under Aelius Gallus opened the Red Sea to Roman use and simultaneously revealed the Arabian Desert as an unsurpassed and, indeed, unsurpassable boundary. The same year Gaius Petronius, the prefect of Egypt, tightened Rome's grip as far as the First Cataract and established a broad military zone beyond it. The vast region north of the Sahara and the Atlas Mountains was also secured (c. 25) after a series of punitive raids against native tribes and the annexation of one client kingdom (Numidia) and the creation of another (Mauretania). Three legions, two in Egypt and one in Africa (a senatorial province), policed the southern shore of the Mediterranean.

In the west, consolidation was extended to the Atlantic. Gaul, Julius Caesar's conquest, was organized as four provinces: senatorial Narbonensis and the imperial three Gauls (Aquitania, Belgica, and Lugdunensis). In Spain, after Agrippa successfully ended in 19 BC the last campaign that Augustus had launched in person in 26, three provinces were formed: senatorial Baetica and imperial Lusitania and Tarraconensis. Three legions enforced Roman authority from Gibraltar to the mouth of the Rhine. Augustus ignored the advice of court poets and others to advance still farther and annex Britain.

In the east, Parthia had demonstrated its power against Crassus and Antony, and Augustus proceeded warily. He retained Antony's ring of buffer client kingdoms, although he incorporated some, including the most celebrated of them, Judaea; he made it a province in AD 6, respecting, however, some of the customs of its Jewish inhabitants. Augustus stationed four legions in Syria and obviously envisaged the Euphrates River and the northern extension of the Arabian Desert as the desirable frontier with Mesopotamia. Farther north, however, no such natural line existed. North of the Black Sea the client kingdom of the Cimmerian Bosporus, under its successive rulers Asander and Polemo, helped to contain southward and westward thrusts by the Scythians, an Iranian people related to the Parthians, and this provided protection in the north for Anatolia and its provinces (senatorial Asia and Bithynia-Pontus and imperial Cilicia and Galatia, the latter a large new province created in 25 BC out of Amyntas' client kingdom). By a show of force, Augustus' stepson Tiberius, in 20 BC, recovered the standards lost at Carrhae and installed Tigranes as client king of Armenia. Although Augustan propaganda depicted this as a famous victory, strategic considerations inevitably obliged the Parthians, once they settled their internal, dynastic dissensions, to dispute Roman control of Armenia. Thus it can hardly be said that Augustus settled the eastern frontier. Missions were sent to the East repeatedly (Agrippa, 17–13 BC; Gaius Caesar, AD 1–4; Germanicus, 18–19), and Armenia remained a problem for Augustus' successors: Tiberius successfully maintained Roman influence there, but Gaius and Claudius failed to do so, leaving Nero with a difficult situation.

Difficulties in the east and north

In the north, too, there was difficulty. The Alps and their passes were finally subjugated early in Augustus' reign. This enabled Tiberius and his brother Drusus between 16 and 8 BC to conquer all the way to the great rivers of central Europe. New provinces were created in the Alps and Tyrol (Maritime and Pennine Alps, Raetia, Noricum) and also farther east (Pannonia, Moesia). Stability along the Danube was precariously maintained, under Augustus and later, by means of periodical alliances with Maroboduus and his successors, who ruled Germanic tribes such as the Marcomanni and Quadi in Bohemia to the north of the river, and by the existence of a Thracian client kingdom to the south of its lowest course. The push across the Rhine began in 12 BC; although it reached the Elbe, consolidation beyond the Rhine proved elusive. A revolt in Pannonia (AD 6–9) interrupted it, and, in AD 9, German tribes under Arminius annihilated Quinctilius Varus and three legions in the Teutoburg Forest. This disaster reduced the number of legions to 25 (it did not reach 28 again until half a century later), and it disheartened Augustus. Old and weary, he withdrew to the Rhine and decided against all further expansion, a policy he urged upon his successor. For the watch on the Rhine the military districts of Upper and Lower Germany were created, containing eight legions between them. Another seven garrisoned the Danubian provinces. These figures reveal imperial anxiety for the northern frontier.

Economic life. Although widespread, Augustus' wars chiefly affected the frontier districts. Elsewhere, peace prevailed. Indeed, never before had so large an area been

Effects of peace

free of war for so long. This state of affairs helped trade. The suppression of piracy and the use of military roads, which the frontier warfare itself brought into being, provided safe arteries of commerce. Stable currency also aided economic growth. Activity directly connected with the soil predominated; but there were also many establishments, usually small, engaged in manufacturing, and such products as textiles, pottery, tiles, and papyrus were turned out in surprising quantities. Advanced techniques were also known: glassblowing, for example, dates from the Augustan age. Most products were consumed locally, but the specialties or monopolies from any region usually exceeded local needs, and the surplus was sold elsewhere, generating a brisk interchange of goods. Some traveled great distances, even beyond the empire: trade with India, for example, reached respectable proportions once the nature of the monsoon was understood, and the Red Sea was opened to Roman shipping. Merchants, especially Levantines, traveled everywhere, and fairs were frequent. The Mediterranean world was linked together as never before, and standardization made considerable headway. In Augustus' day Italy was economically the most important part of the empire. It could afford to import on a large scale, thanks partly to provincial tribute but above all to its own large productivity. The eastern provinces, for their part, recovered rapidly from the depredations of the civil wars and were industrially quite advanced. The other provinces were less developed, but they soon ceased being mere suppliers of raw materials; they learned to exploit their natural resources by using new techniques and then began overtaking the more advanced economies of Italy and the Greek-speaking regions. The importance of trade in unifying the empire should not be underestimated.

Augustan art and literature. In 17 BC Rome held Secular Games, a traditional celebration to announce the entry into a new epoch (*saeculum*). New it was, for, though Augustus preserved what he could of republican institutions, he added much that was his own. His Rome had become very Italian, and this spirit is reflected in the art and literature of his reign. Its greatest writers were native Italians, and, like the ruler whose program they glorified, they used the traditional as the basis for something new. Virgil, Horace, and Livy, as noted above, imitated the writing of classical Greece, but chiefly in form, their tone and outlook being un-Hellenic. It was the glory of Italy and faith in Rome that inspired Virgil's *Georgics* and *Aeneid,* Horace's *Odes,* and the first 10 books of Livy's history.

In Augustan art a similar fusion was achieved between the prevailing Attic and Hellenistic models and Italian naturalism. The sculptured portraits on the Ara Pacis Augustae (Altar of the Augustan Peace) of 9 BC, for all their lifelike quality, are yet in harmony with the classical poise of the figures, and they strike a fresh note: the stately converging processions (Rome's imperial family and magistrates on one side; senators, equites, and citizens on the other) became the prototypes for all later processional reliefs. Augustan painting likewise displays a successful combination of Greek and Roman elements, to judge from the frescoes in the house of Livia on the Palatine. In Augustan architecture, decidedly conservative and Hellenic, the potentialities of curving and vaulted spaces that had been revealed in the earlier 1st century BC were not realized. Building was, however, very active and widespread.

Cultural excellence of the Augustan age

The culture of the age undoubtedly attained a high level of excellence, dominated by the personality of the emperor and his accomplishments. Imperial art had already reached full development, a matter of no small moment, because Rome's political predominance made the spread of its influence inevitable. The Mediterranean world was soon assuming a Roman aspect, and this is a measure of Augustus' extraordinary achievement. Yet it was an achievement with limitations. His professed aim—to promote stability, peace, security, and prosperity—was irreproachable, but perhaps it was also unexciting. Emphasizing conservatism by precept and his own example, he encouraged the simpler virtues of a less sophisticated age, and his success made this sedate but rather static outlook fashionable. People accepted the routine of his continuing rule, at the cost, however, of some loss of intellectual energy and moral fervour. The great literature, significantly, belongs to the years near Actium, when people's imagination still nursed heady visions of Roman victory and Italian destiny. After the Secular Games the atmosphere became more commonplace and produced the frivolities of Ovid and the pedestrian later books of Livy.

Appraisal of Augustus. Augustus' position as princeps cannot be defined simply. He was neither a Roman king (rex) nor a Hellenistic monarch (*basileus*), nor was he, as the 19th-century German historian Theodor Mommsen thought, a partner with the Senate in a dyarchy. He posed as the first servant of an empire over which the Roman Senate presided, and it would appear that his claim to have accepted no office inconsistent with ancestral custom was literally true. Proconsular *imperium* was a republican institution, and, although tribunician power was not, it contained nothing specifically unrepublican. But, while precedents can be cited for Augustus' various powers, their concentration and tenure were absolutely unparalleled. Under the republic, powers like his would have been distributed among several holders, each serving for a limited period with a colleague. Augustus wielded them all, by himself, simultaneously and without any time limit (in practice, at least). This fact made him an emperor, but it did not necessarily make him a military tyrant.

In discharging both military and civilian functions, Augustus was no different from republican consuls or praetors. Admittedly his military power was overwhelming; but, if he chose not to brandish it, the tone of his reign could remain essentially civilian. Constitutional safeguards were indeed lacking; everything was at the emperor's discretion, and even Augustus passed legislation that made anti-imperial behaviour, real or suspected, treasonable (men were, in fact, executed for conspiracy during his reign). But there had been no constitutional safeguards in the republic, under Sulla, Pompey, the triumvirs, or even Julius Caesar. Augustus' improved police services probably made lower-class Romans at least feel safer under him. The senatorial class, however, contained a minority resentful of the sheer undeniable preponderance of the princeps' power, and he was the target of several unsuccessful plots against his life.

The principate was something personal, what the emperor chose to make it, and the relations prevailing between emperor and Senate usually indicated what a reign was like. In Augustus' case they reveal a regime that was outwardly constitutional, generally moderate, and certainly effective. But, as he himself implied at the end of his life, he was a skillful actor in life's comedy. Later emperors lacked his sureness of touch.

Augustus' accomplishments

When Augustus died, the Senate unhesitatingly pronounced him *divus*—the deified one who had restored peace, organized a standing army to defend the frontiers, expanded those frontiers farther than any previous Roman, improved administrative practices everywhere, promoted better standards of public and private behaviour, integrated Rome and Italy, embellished Rome, reconciled the provinces, expedited Romanization, and above all maintained law and order while respecting republican traditions.

Augustus' luck was hardly inferior to his statecraft. Despite indifferent health, he headed the Roman state in one capacity or another for 56 years. His rule, one of the longest in European history, consolidated the principate so firmly that what might have been an episode became an epoch. At his death there was practically no one left with any personal memory of the republic, and Augustus' wish came true: he had fashioned a lasting as well as constitutional government. The principate endured with only minor changes for about 200 years.

The succession. Like any great Roman magnate, Augustus owed it to his supporters and dependents to maintain the structure of power which they constituted together and which would normally pass from father to son. In accepting the heritage from Caesar, he had only done the right thing, and he was respected for it by his peers. None of them would have advised him later to dismantle what he had since added to it. When, for instance, he was away from Rome, rather than accepting a diminution in his pre-

rogatives of administration, a senator as city prefect was deputed to represent him. Consequently, Augustus began thinking early about who should follow him. The soldiers' views on legitimacy reinforced his own natural desire to found a dynasty, but he had no son and was therefore obliged to select his successor. Death played havoc with his attempts to do so. His nephew Marcellus, his son-in-law Agrippa, his grandsons Gaius and Lucius (Julia's children by Agrippa), were groomed in turn; but they all predeceased him. Augustus, finally and reluctantly, chose a member of the republican nobility, his stepson Tiberius, a scion of the ultra-aristocratic Claudii. In AD 4 Augustus adopted Tiberius as his son and had tribunician power and probably proconsular *imperium* as well conferred upon him. This arrangement was confirmed in 13, and, when Augustus died the following year, Tiberius automatically became emperor.

Tiberius (ruled 14–37), during whose reign Christ was crucified, was a soldier and administrator of proved capability but of a reserved and moody temperament that engendered misunderstanding and unpopularity. Slander blamed him for the death in 19 of his nephew and heir apparent, the popular Germanicus; and, when informers (*delatores*), who functioned at Rome like public prosecutors, charged notables with treason, Tiberius was thought to encourage them. By concentrating the praetorian cohorts in a camp adjoining Rome, he increased the soldiers' scope for mischief-making without building any real security, and in 26 he left Rome permanently for the island of Capreae (Capri), entrusting Rome to the care of the city prefect. Tiberius heeded the aged Augustus' advice and did not extend the empire. (The annexation of Cappadocia, a client kingdom, represented no departure from Augustan policy.) In general he took his duties seriously; however, by administering the empire from Capreae he offended the Senate and was never fully trusted, much less really liked. At his death he was not pronounced *divus*. His great-nephew, Germanicus' son Gaius, succeeded him.

Gaius (better known by his nickname, Caligula, meaning Little Boot) ruled from 37 to 41 with the absolutism of an Oriental monarch: his short reign was filled with reckless spending, callous murders, and humiliation of the Senate. Gaius' foreign policy was inept. Projected annexation proved abortive in Britain; it touched off heavy fighting in Mauretania. In Judaea and Alexandria, Gaius' contemptuous disregard of Jewish sentiment provoked near rebellion. When assassination ended his tyranny, the Senate contemplated restoration of the republic but was obliged by the Praetorian Guard to recognize Claudius, Germanicus' brother and therefore Gaius' uncle, as emperor.

Centralization and expansion under Claudius I

Claudius I (ruled 41–54) went far beyond Augustus and Tiberius in centralizing government administration and, particularly, state finances in the imperial household. His freedmen secretaries consequently acquired great power; they were in effect directors of government bureaus. Claudius himself displayed much interest in the empire overseas; he enlarged it significantly, incorporating client kingdoms (Mauretania in 42; Lycia, 43; Thrace, 46) and, more important, annexing Britain. Conquest of Britain began in 43, Claudius himself participating in the campaign; the southeast was soon overrun, a *colonia* established at Camulodunum (Colchester) and a *municipium* at Verulamium (St. Albans), while Londinium (London) burgeoned into an important entrepôt. Claudius also promoted Romanization, especially in the western provinces, by liberally granting Roman citizenship, by founding *coloniae*, and by inducting provincials directly into the Senate—he became censor in 47 and added to the Senate men he wanted, bestowing appropriate quaestorian or praetorian rank upon them to spare the maturer ones among them the necessity of holding junior magistracies; lest existing senators take offense, he elevated some of them to patrician status (a form of patronage often used by later emperors). Claudius' provincial policies made the primacy of Italy less pronounced, although that was hardly his aim. In fact, he did much for Italy, improving its harbours, roads, and municipal administration and draining its marshy districts. The execution of many senators and equites, the insolence and venality of his freedmen, the excessive influence of his wives, and even his bodily infirmities combined to make him unpopular. Nevertheless, when he died (murdered probably by his fourth wife, Agrippina, Augustus' great-granddaughter, who was impatient for the succession of the 16-year-old Nero, her son by an earlier marriage), he was pronounced *divus*.

The reign of Nero

Nero (ruled 54–68) left administration to capable advisers for a few years but then asserted himself as a vicious despot. He murdered successively his stepbrother Britannicus, his mother Agrippina, his wife Octavia, and his tutor Seneca. He also executed many Christians, accusing them of starting the great fire of Rome in 64 (this is the first recorded Christian persecution). In Rome his reliance on Oriental favourites and his general misgovernment led to a conspiracy by Gaius Calpurnius Piso in 65, but it was suppressed, leading to yet more executions; the victims included the poet Lucan. The empire was not enlarged under this unwarlike emperor, but it was called upon to put down serious disorders. In Britain in 60–61 the rapacity and brutality of Roman officials provoked a furious uprising under Queen Boudicca; thousands were slaughtered, and Camulodunum, Vernulamium, and Londinium were destroyed. In the east a major military effort under Corbulo, Rome's foremost general, was required (62–65) to reestablish Roman prestige; a compromise settlement was reached, with the Romans accepting the Parthian nominee in Armenia and the Parthians recognizing him as Rome's client king. In 66, however, revolt flared in Judaea, fired by Roman cruelty and stupidity, Jewish fanaticism, and communal hatreds; the prefect of Egypt, Julius Alexander, prevented involvement of the Jews of the Diaspora. An army was sent to Judaea under Titus Flavius Vespasianus to restore order; but it had not completed its task when two provincial governors in the west rebelled against Nero—Julius Vindex in Gallia Lugdunensis and Sulpicius Galba in Hispania Tarraconensis. When the praetorians in Rome also renounced their allegiance, Nero lost his nerve and committed suicide. He brought the Julio-Claudian dynasty to an ignominious end by being the first emperor to suffer *damnatio memoriae*—his reign was officially stricken from the record by order of the Senate.

Growth of the empire under the Flavians and Antonines. *The year of the four emperors.* Nero's death ushered in the so-called year of the four emperors. The extinction of the Julio-Claudian imperial house robbed the soldiers of a focus for their allegiance, and civil war between the different armies ensued. The army of Upper Germany, after crushing Vindex, urged its commander, Verginius Rufus, to seize the purple for himself. But he elected to support Galba—scion of a republican patrician family claiming descent from Jupiter and Pasiphae—who was recognized as emperor by the Senate. However, the treasury, emptied by Nero's extravagance, imposed a stringent economy, and this bred unpopularity for Galba; his age (73) was also against him, and unrest grew. Early in January 69 the Rhineland armies acclaimed Aulus Vitellius, commander in Lower Germany; at Rome the praetorians preferred Marcus Salvius Otho, whom Galba had alienated by choosing a descendant of the old republican aristocracy for his successor. Otho promptly procured Galba's murder and obtained senatorial recognition; this ended the monopoly of the purple for the republican nobility.

Otho, however, lasted only three months; defeated at Bedriacum, near Cremona in northern Italy, by Vitellius' powerful Rhineland army, he committed suicide (April 69). The Senate thereupon recognized Vitellius; but the soldiers along the Danube and in the east supported Vespasianus, the commander in Judaea. In a second battle near Bedriacum, the Rhineland troops were defeated in their turn, and on Vitellius' death soon afterward an accommodating Senate pronounced Vespasian emperor.

The Flavian emperors. On Dec. 22, 69, the Senate conferred all the imperial powers upon Vespasian *en bloc* with the famous Lex de Imperio Vespasiani ("Law Regulating Vespasian's authority"), and the Assembly ratified the Senate's action. This apparently was the first time that such a law was passed; a fragmentary copy of it is preserved on the Capitol in Rome.

Vespasian (ruled 69–79) did not originate from Rome or

its aristocracy. His family came from the Sabine municipality Reate, and with his elevation the Italian bourgeoisie came into its own. He and his two sons, both of whom in turn succeeded him, constituted the Flavian dynasty (69–96). Vespasian faced the same difficult task as Augustus—the restoration of peace and stability. The disorders of 69 had taken troops away from the Rhine and Danube frontiers. Thereupon, the Danubian lands were raided by Sarmatians, a combination of tribes who had overwhelmed and replaced the Scythians, their distant kinsmen, in eastern Europe. The assailants were repelled without undue difficulty; but the Sarmatian Iazyges, now firmly in control of the region between the Tisza and Danube rivers, posed a threat for the future.

Developments in the Rhineland were more immediately serious. There in 69 a certain Civilis incited the Batavians serving as auxiliaries in the Roman army to rebel. Gallic tribes joined the movement, and the insurgents boldly overran all but two of the legionary camps along the Rhine. Vespasian sent his relative Petilius Cerealis to deal with the rebels, who, fortunately for Rome, were not united in their aims; by 70 Cerealis had restored order. That same year Vespasian's elder son, Titus, brought the bloody war in Judaea to its end by besieging, capturing, and destroying Jerusalem.

Vespasian's policies

To rehabilitate the public finances, Vespasian introduced new imposts, including a poll tax on Jews, and practiced stringent economies. With the Senate he was courteous but firm. He allowed it little initiative but used it as a reservoir from which to obtain capable administrators. To that end he assumed the censorship and added senators on a larger scale than Claudius had done, especially from the municipalities of Italy and the western provinces. Already before 69 an aristocracy of service had arisen, and the provincialization of the Roman Senate had begun; thereafter this development made rapid headway. Besides the censorship, Vespasian also often held the consulship, usually with Titus as his colleague. His object presumably was to ensure that his own parvenu Flavian house outranked any other. In this he succeeded; the troops especially were ready to accept the Flavians as the new imperial family. On Vespasian's death in 79, Titus, long groomed for the succession, became emperor and immediately had his father deified.

Titus (ruled 79–81) had a brief reign, marred by disasters (the volcanic eruption that buried Pompeii and Herculaneum and another great fire in Rome); but his attempts to alleviate the suffering and his general openhandedness won him such popularity that he was unhesitatingly deified after his early death.

Domitian (ruled 81–96), Titus' younger brother, had never been formally indicated for the succession; but the praetorians acclaimed him, and the Senate ratified their choice. Throughout his reign Domitian aimed at administrative efficiency, but his methods were high-handed. For him the Senate existed merely to supply imperial servants. He also used equites extensively, more than any previous emperor. He held the consulship repeatedly, was *censor perpetuus* from 85 on, and demanded other extravagant honours. On the whole, his efficiency promoted the welfare of the empire. Above all, he retained the allegiance of the troops. Although scornful of the Senate's dignity, he insisted on his own and mercilessly punished any act of disrespect, real or fancied, toward himself. He became even more suspicious and ruthless when Saturninus, commander in Upper Germany, attempted rebellion in 89. He crushed Saturninus; executions and confiscations ensued, and *delatores* flourished. The tyranny was particularly dangerous to senators, and it ended only with Domitian's assassination in 96. The Flavian dynasty, like the Julio-Claudian, ended with an emperor whose memory was officially damned.

Military reforms under the Flavians

The disorders in 69 were the cause of some military reforms. Under the Flavians, auxiliaries usually served far from their native hearths under officers of different nationality from themselves. At the same time, the tasks assigned to them came increasingly to resemble those performed by the legionaries. The latter grew less mobile, as camps with stone buildings came to be the rule; and it became common for detachments from a legion (*vexillationes*), rather than the entire legion, to be used for field operations. This army of a new type proved its mettle in Britain, where the advance halted by Boudicca's revolt was now resumed. Between 71 and 84 three able governors—Petilius Cerealis, Julius Frontinus, and Julius Agricola, the latter Tacitus' father-in-law—enlarged the province to include Wales and northern England; Agricola even reached the Scottish highlands before Domitian recalled him.

Along the Rhine, weaknesses revealed by Civilis' revolt were repaired. Vespasian crossed the river in 74 and annexed the Agri Decumates, the triangle of land between the Rhine, Danube, and Main rivers. To consolidate the position, he and Domitian after him penetrated the Neckar River valley and Taunus mountains, and fortifications began to take shape to the east of the Rhine, a military boundary complete with strongpoints, watchtowers, and, later, a continuous rampart of earthworks and palisades. Once Saturninus' revolt in 89 had been suppressed, Domitian felt the situation along the Rhine sufficiently stable to warrant conversion of the military districts of Upper and Lower Germany into regular provinces and the transfer of some Rhineland troops to the Danube. To the north of this latter river, the Dacians had been organized into a strong kingdom, ruled by Decebalus and centring on modern Romania; in 85 they raided southward across the Danube, and in the next year they defeated the Roman punitive expedition. Domitian restored the situation in 88, but Saturninus' rebellion prevented him from following up his success. Domitian and Decebalus thereupon came to terms: Decebalus was to protect the lower Danube against Sarmatian attack, and Domitian was to pay him an annual subsidy in recompense. The Danubian frontier, however, remained disturbed, and Domitian wisely strengthened its garrisons; by the end of his reign it contained nine legions, as against the Rhineland's six, and Pannonia was soon to become the military centre of gravity of the empire.

Strengthening of the eastern frontier

The Flavians also took measures to strengthen the eastern frontier. In Asia Minor, Vespasian created a large "armed" province by amalgamating Cappadocia, Lesser Armenia, and Galatia; and the whole area was provided with a network of military roads. South of Asia Minor, Judaea was converted into an "armed" province by getting legionary troops; and two client kingdoms—Commagene and Transjordan—were annexed and added to Syria. Furthermore, the legionary camps seem now to have been established right on the Euphrates at the principal river crossings. This display of military strength kept the empire and Parthia at peace for many years.

The Antonine emperors. Marcus Cocceius Nerva, an elderly senator of some distinction, was the choice of Domitian's assassins for emperor; and the Senate promptly recognized him. The soldiers, however, did so much more reluctantly, and, because the year 69 had revealed that emperors no longer needed to be Roman aristocrats and could be chosen in places other than Rome, their attitude imposed caution.

Nerva, who ruled from 96 to 98, adopted a generally lavish and liberal policy, but it failed to win the soldiers over completely, and he proved unable to save all Domitian's murderers from their vengeance. Unrest subsided only when, overlooking kinsmen of his own, he adopted an outstanding soldier, Marcus Ulpius Trajanus, who was governor of Upper Germany, as his successor. Nerva himself died a few months later.

Trajan (ruled 98–117) was the first and perhaps the only emperor to be adopted by a predecessor totally unrelated to him by either birth or marriage. He was also the first in a series of "good" rulers who succeeded one another by adoption and for most of the 2nd century provided the empire with internal harmony and careful government; they are collectively, if somewhat loosely, called the Antonine emperors. More significantly still, Trajan, a Spaniard, was also the first princeps to come from the provinces; with the greater number of provincials now in the Senate, the elevation of one of them, sooner or later, was practically inevitable. Throughout his reign, Trajan generally observed constitutional practices. Mindful of the susceptibilities of the Senate, he regularly consulted and

reported to it. Modest in his bearing, he did not claim ostentatious honours such as frequent consulships or numerous imperial salutations, and he mixed easily with senators on terms of cordial friendship. This reestablished mutual respect between princeps and Senate. Empire and liberty, in Tacitus' words, were reconciled, and the atmosphere of suspicion, intrigue, and terror surrounding the court in Domitian's day disappeared. Trajan endeared himself also to the populace at large with lavish building programs, gladiatorial games, and public distributions of money. Above all, he was popular with the armed forces; he was the soldier-emperor par excellence. Understandably, he received the title Optimus (Best), officially from 114 on (and unofficially for many years earlier).

Yet Trajan was a thoroughgoing autocrat who intervened without hesitation or scruple even in the senatorial sphere, whenever it seemed necessary. His aim was efficiency; his desire was to promote public welfare everywhere. He embellished Rome with splendid and substantial structures, and he showed his care for Italy by refurbishing and enlarging the harbours at Ostia, Centumcellae, and Ancona. He sent officials called *curatores* to Italian municipalities in financial difficulties and helped to rehabilitate them. He greatly expanded an ingenious charity scheme probably begun by Nerva: money was loaned to farmers on easy terms, and the low interest they paid went into a special fund for supporting indigent children. Nor did Trajan neglect Italy's highway network: he built a new road (Via Traiana) that soon replaced the Via Appia as the main thoroughfare between Beneventum and Brundisium.

Interest in Italy implied no neglect of the provinces. *Curatores* were also sent to them; to rescue Achaea and Bithynia, senatorial provinces, from threatened bankruptcy, Trajan made them both temporarily imperial, sending special commissioners of his own to them. His correspondence with his appointee in Bithynia, the younger Pliny, has survived and reveals how conscientiously the emperor responded on even the smallest details. At the same time, it reveals how limited was access to the central government and, consequently, how great a latitude for independent decisions must be left to the governors who lacked some special claim on the emperor's attention. Trajan's day was too short to hear every speech of every delegation from the provinces, every recommendation to bestow favour or grant promotion, and every appeal to himself as supreme judiciary. To assist him, he had a "bureaucracy" of only a few hundred in Rome and a few more hundred serving in various capacities in the provinces—to direct the lives of some 60 million people. Clearly, most government must in fact rest in the hands of local aristocracies.

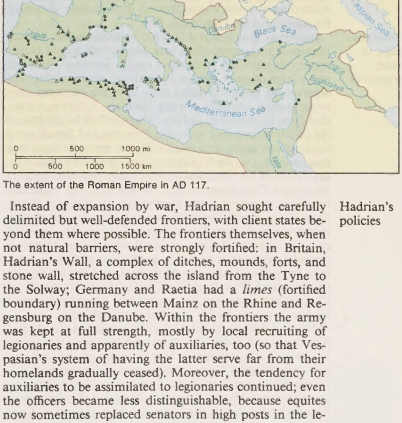
Expansion of the empire under Trajan

In the military sphere, Trajan's reign proved a most dynamic one. He decided to strengthen the dangerous Danube frontier by converting Dacia into a salient of Roman territory north of the river in order to dismember the Sarmatian tribes and remove the risk of large, hostile combinations to a safer distance. Bringing to bear a force of 100,000 men, he conquered Decebalus in two hard-fought wars (101–102; 105–106) and annexed Dacia, settling it with people from neighbouring parts of the empire. On the eastern frontier he planned a similar operation, evidently in the conviction, shared by many eminent Romans both before and after him, that only conquest could solve the Parthian problem. Possibly, too, he wished to contain the menace of the Sarmatian Alani in the Caspian region. In a preliminary move, the Nabataean kingdom of Arabia Petraea was annexed in 105–106. Then, in 114, Trajan assembled another large army, incorporated the client kingdom of Armenia, and invaded Parthia.

After spectacular victories in 115 and 116, he created additional provinces (Northern Mesopotamia, Assyria) and reached the Persian Gulf. But he had merely overrun Mesopotamia; he had not consolidated it, and, as his army passed, revolts broke out in its rear. The Jews of the Diaspora and others seized their chance to rebel, and before the end of 116 much of the Middle East besides Parthia was in arms (Cyrene, Egypt, Cyprus, Anatolia). Trajan proceeded resolutely to restore the situation, but death found him still in the East.

Before his last illness he had not formally indicated his successor. But high honours and important posts had been accorded his nearest male relative, Publius Aelius Hadrianus, the governor of Syria; and, according to Trajan's widow, Hadrian had actually been adopted by Trajan on his deathbed. Accordingly, both Senate and soldiers recognized him. Trajan's posthumous deification was never in doubt.

Hadrian (ruled 117–138), also a Spaniard, was an emperor of unusual versatility. Unlike Trajan, he was opposed to territorial expansion. Being himself in the East in 117, he renounced Trajan's conquests there immediately and contemplated evacuating Dacia as well. Furthermore, four of the consular generals particularly identified with Trajan's military ventures were arrested and executed "for conspiracy"; Hadrian claimed later that the Senate ordered their deaths against his wishes. The only heavy fighting during his generally peaceful reign occurred in Judaea—or Syria Palaestina, as it was thenceforth called—where Bar Kokhba led a furious, if futile, Jewish revolt (132–135) against Hadrian's conversion of Jerusalem into a Roman colony named Aelia Capitolina.



The extent of the Roman Empire in AD 117.

Hadrian's policies

Instead of expansion by war, Hadrian sought carefully delimited but well-defended frontiers, with client states beyond them where possible. The frontiers themselves, when not natural barriers, were strongly fortified: in Britain, Hadrian's Wall, a complex of ditches, mounds, forts, and stone wall, stretched across the island from the Tyne to the Solway; Germany and Raetia had a *limes* (fortified boundary) running between Mainz on the Rhine and Regensburg on the Danube. Within the frontiers the army was kept at full strength, mostly by local recruiting of legionaries and apparently of auxiliaries, too (so that Vespasian's system of having the latter serve far from their homelands gradually ceased). Moreover, the tendency for auxiliaries to be assimilated to legionaries continued; even the officers became less distinguishable, because equites now sometimes replaced senators in high posts in the legions. To keep his essentially sedentary army in constant readiness and at peak efficiency (no easy task), Hadrian carried out frequent personal inspections, spending about half his reign in the provinces (121–125; 128–134).

Hadrian also was responsible for significant developments on the civilian side. Under him, equites were no longer required to do military service as an essential step in their career, and many of them were employed in the imperial civil service, more even than under Domitian. By now the formative days of the civil service were over; its bureaucratic phase was beginning, and it offered those equites who had no military aspirations an attractive, purely civilian career. Formal titles now marked the different equestrian grades of dignity: a procurator was *vir egregius;* an ordinary prefect, *vir perfectissimus;* a praetorian prefect, *vir eminentissimus,* the latter title being obviously parallel to the designation *vir clarissimus* for a senator. Thenceforth, equites replaced freedmen in the

imperial household and bureaus, and they even appeared in Hadrian's imperial council.

Hadrian also improved legal administration. He had his expert jurists codify the *edictum perpetuum* (the set of rules gradually elaborated by the praetors for the interpretation of the law). He also appointed four former consuls to serve as circuit judges in Italy. This brought Italy close to becoming a province; Hadrian's intent, however, was not to reduce the status of Italy but to make all parts of the empire important. For one part of his realm, he was exceptionally solicitous: he spent much time in Greece and lavishly embellished Athens.

Hadrian maintained good relations with but was never fully trusted by the Senate. His foreign policy seemed to be unheroic, his cosmopolitanism to be un-Roman, and his reforms to encroach on activities traditionally reserved for senators. Moreover, in his last two years he was sometimes capricious and tyrannous. Like Augustus, he had no son of his own and conducted a frustrating search for a successor. After executing his only male blood relative, his grandnephew, in 136, he adopted Lucius Ceionius Commodus, renaming him Lucius Aelius Caesar. The latter, however, died shortly afterward, whereupon Hadrian in 138 chose a wealthy but sonless senator, the 51-year-old Titus Aurelius Antoninus; but, evidently intent on founding a dynasty, he made Antoninus in his turn adopt two youths, 16 and 7 years old, respectively—they are known to history as Marcus Aurelius (the nephew of Antoninus' wife) and Lucius Verus (the son of Aelius Caesar). When Hadrian died soon thereafter, Antoninus succeeded and induced a reluctant Senate to deify the deceased emperor. According to some, it was this act of filial piety that won for Antoninus his cognomen, Pius.

Antoninus' reign of quiet prosperity

Antoninus Pius (ruled 138–161) epitomizes the Roman Empire at its cosmopolitan best. He himself was of Gallic origin; his wife was of Spanish origin. For most men his was a reign of quiet prosperity, and the empire under him deserves the praises lavished upon it by the contemporary writer Aelius Aristides. Unlike Hadrian, Antoninus traveled little; he remained in Italy, where in 148 he celebrated the 900th anniversary of Rome. Princeps and Senate were on excellent terms, and coins with the words *tranquillitas* and *concordia* on them in Antoninus' case mean what they say. Other of his coins not unreasonably proclaim *felicitas temporum* ("the happiness of the times"). Yet raids and rebellions in many of the borderlands (in Britain, Dacia, Mauretania, Egypt, Palaestina, and elsewhere) were danger symptoms, even though to the empire at large they seemed only faraway bad dreams, to use the expression of Aelius Aristides. Antoninus prudently pushed the Hadrianic frontiers forward in Dacia, the Rhineland, and Britain (where the Antonine Wall from the Firth of Forth to the River Clyde became the new boundary) and carefully groomed his heir apparent for his imperial responsibilities.

Marcus Aurelius (ruled 161–180) succeeded the deified Antoninus and more than honoured Hadrian's intentions by immediately co-opting Lucius Verus as his full coemperor. Because Verus' competence was unproved, this excess of zeal was imprudent. Fortunately, Verus left decision making to Marcus. Marcus' action was also dangerous for another reason; it represented a long step away from imperial unity and portended the ultimate division of the empire into Greek- and Latin-speaking halves. Nor was this the only foreboding development in Marcus' reign—formidable barbarian assaults were launched against the frontiers, anticipating those that were later to bring about the disintegration of the empire. Marcus himself was a stoic philosopher; his humanistic, if somewhat pessimistic, *Meditations* reveal how conscientiously he took his duties. Duty called him to war; he responded to the call and spent far more of his reign in the field than had any previous emperor.

At Marcus' very accession the Parthians turned aggressive, and he sent Verus to defend Roman interests (162). Verus greedily took credit for any victories but left serious fighting to Avidius Cassius and the army of Syria. Cassius succeeded in overrunning Mesopotamia and even took Ctesiphon, the Parthian capital; he was therefore able to conclude a peace that safeguarded Rome's eastern provinces and client kingdoms (166). In the process, however, his troops became infected with plague, and they carried it back with them to the west with calamitous results. The Danube frontier, already weakened by the dispatch of large detachments to the East, collapsed under barbarian assault. Pressed on from behind by Goths, Vandals, Lombards, and others, the Germanic Marcomanni and Quadi and the Sarmatian Iazyges poured over the river; the Germans actually crossed Raetia, Noricum, and Pannonia to raid northern Italy and besiege Aquileia. Marcus and Verus relieved the city shortly before Verus' death (169). Then, making Pannonia his pivot of maneuver, Marcus pushed the invaders back; by 175 they were again beyond the Danube. At that moment, however, a false report of Marcus' death prompted Avidius Cassius, by now in charge of all eastern provinces, to proclaim himself emperor. The news of this challenge undid Marcus' achievements along the Danube because it took him to the East and reopened the door to barbarian attacks. Fortunately, Cassius was soon murdered, and Marcus could return to central Europe (177). But he had barely restored the frontier again when he died at Vindobona (Vienna) in 180, bequeathing the empire to his son, the 19-year-old Commodus, who had actually been named coemperor three years earlier.

Barbarian assaults

Commodus (ruled 180–192), like Gaius and Nero, the youthful emperors before him, proved incompetent, conceited, and capricious. Fortunately, the frontiers remained intact, thanks to able provincial governors and to barbarian allies, who had been settled along the Danube with land grants and who gave military service in return. But Commodus abandoned Marcus' scheme for new trans-Danubian provinces, preferring to devote himself to sensual pleasures and especially to the excitements of the arena in Rome, where he posed as Hercules Romanus and forced the Senate to recognize his godhead officially. He left serious business to his favourites, whose ambitions and intrigues led to plots, treason trials, confiscations, and insensate murders. Commodus' assassination on the last day of 192 terminated a disastrous reign; thus the Antonines, like the Julio-Claudians, had come to an ignominious end. And there was a similar sequel. Commodus' *damnatio memoriae,* like Nero's, was followed by a year of four emperors.

The empire in the 2nd century. The century and three-quarters after Augustus' death brought no fundamental changes to the principate, although so long a lapse of time naturally introduced modifications and shifts of emphasis. By Flavian and Antonine times the principate was accepted universally. For the provinces, a return to the republic was utterly unthinkable; for Rome and Italy, the year 69 served as a grim warning of the chaos to be expected if, in the absence of a princeps, the ambitions of a few powerful individuals obtained unfettered scope. A princeps was clearly a necessity, and people were even prepared to tolerate a bad one, although naturally they always hoped for a good one.

The princeps, moreover, did not have to be chosen any longer from the Julio-Claudians. The great achievement of the Flavians was to reconcile the soldiers and the upper classes everywhere to the idea that others were eligible. The Flavians' frequent tenure of consulship and censorship invested their family, though not of the highest nobility, with the outward trappings of prestige and the aristocratic appearance of an authentic imperial household. The deification of the first two Flavians contributed to the same end, and so did the disappearance of old republican families that might have outranked the reigning house (by 69 most descendants of the republican nobility had either died of natural causes or been exterminated by imperial persecution). After the Flavians, the newness of a man's senatorial dignity and the obscurity of his ultimate origin, whether it was Italian or otherwise, no longer forbade his possible elevation. Indeed, Domitian's successors and even Domitian himself in his last years did not need to enhance their own importance by repeated consulships. The Antonine emperors, like the Julio-Claudians, held the office infrequently. They did, however, continue the Flavian practice of emphasizing the loftiness of their families

Eligibility for the principate

by deifying deceased relatives (Trajan deified his sister, his niece, and his father; Antoninus, his wife; and so forth).

Trend to absolute monarchy. Glorification of the reigning house, together with a document such as Vespasian's Lex de Imperio, helped to advertise the emperor's position; and under the Flavians and Antonines the principate became much more like an avowed monarchy. Proconsular *imperium* began to be reflected in the imperial titulary, and official documents started calling the emperor *dominus noster* ("our master").

The development of imperial law-making clearly illustrates the change. From the beginnings of the principate, the emperor had had the power to legislate, although no law is known that formally recognized his right to do so; by Antonine times, legal textbooks stated unequivocally that whatever the emperor ordered was legally binding. The early emperors usually made the Senate their mouthpiece and issued their laws in the form of senatorial decrees; by the 2nd century the emperor was openly replacing whatever other sources of written law had hitherto been permitted to function. After 100 the Assembly never met formally to pass a law, and the Senate often no longer bothered to couch its decrees in legal language, being content to repeat verbatim the speech with which the ruler had advocated the measure in question. After Hadrian, magistrates ceased modifying existing law by their legal interpretations because the praetors' *edictum perpetuum* had become a permanent code, which the emperor alone could alter. By 200, learned jurists had lost the right they had enjoyed since the time of Augustus of giving authoritative rulings on disputed points (*responsa prudentium*). Meanwhile, the emperor more and more was legislating directly by means of edicts, judgments, mandates, and rescripts—collectively known as *constitutiones principum.* He usually issued such *constitutiones* only after consulting the "friends" (*amici Caesaris*) who composed his imperial council. But a *constitutio* was nevertheless a fiat. The road to the later dominate (after 284) lay open.

Political life. Nevertheless, the autocratic aspect of the Flavian and Antonine regimes should not be overstressed. Augustus himself had been well aware that it was impossible to disguise permanently the supremacy that accumulation of powers gained piecemeal conferred; his deportment in his last years differed little from that of Vespasian, Titus, and the so-called five good emperors who followed them. Nor had other Julio-Claudians hesitated to parade their predominance—Claudius, by centralizing the imperial powers, reduced their apparent diversity to one all-embracing *imperium;* Gaius and Nero revealed the autocracy implicit in the principate with frank brutality.

Civilian character of the Flavian and Antonine reigns

What impresses perhaps as much as the undoubtedly autocratic behaviour of the Flavians and Antonines is the markedly civilian character of their reigns. They held supreme power, and some of them were distinguished soldiers; yet they were not military despots. For this the old republican tradition—whereby a state official might serve in both a civilian and a military capacity—was largely responsible. Matters, however, were open to change after Hadrian separated the two realms of service. Actually, the 3rd century soon showed what it meant to have a princeps whose whole experience had been confined to camps and barracks.

As imperial powers became more concentrated, republican institutions decayed; the importance of imperial officials grew, while the authority of urban magistrates declined. Quaestorship, praetorship, and consulship (the last-named now reduced to a two-month sinecure) became mere stepping stones to the great imperial posts that counted most in the life of the empire. Governors of imperial provinces and commanders of legions were Roman senators; but they were equally imperial appointees. Clearly, the emperor was the master of the Senate; and it was disingenuous for him to get impatient, as some emperors did, with the Senate's lack of initiative and reluctance to take firm decisions of its own. The emperor might not even consult the Senate much, preferring to rely on his imperial council, in which equestrian bureau chiefs over the course of the 2nd century came to constitute an established element.

The Senate, however, at least until the reign of Commodus, was treated courteously by most Flavians and Antonines. They recognized its importance as a lawcourt, as the body that formally appointed a new emperor, and as a sounding board of informed opinion. Senators came increasingly from the provinces, and, although this meant preeminently the western provinces (the Greek-speaking East being underrepresented), the Senate did reflect to some extent the views of the empire at large.

The equites, meanwhile, steadily acquired greater importance as imperial officials. In newly created posts they invariably became the incumbents, and in posts of long standing they replaced freedmen and *publicani.* During the 2nd century equestrian procurators increased markedly in numbers as the direction of imperial business came to be more tidily subdivided. Four grades of service distinguished by salary were established. While the government assumed a more rational flow and outline, its total number of employees nevertheless remained quite tiny, compared with that of the 4th and later centuries.

Rome and Italy. By the 2nd century the city of Rome had attracted freeborn migrants from all over the empire; it housed, additionally, large numbers of manumitted slaves. These newcomers were all assimilated and diluted the city's Italian flavour. The vast majority of them were poor, the handful of opulent imperial freedmen being entirely exceptional. But many were energetic, enterprising, and lucky, able to make their way in the world. Freedmen laboured under a social stigma, although some of them managed to become equites. Their sons, however, might overcome discrimination, and their grandsons were even eligible for membership in the Senate.

The urban plebs

Inevitably, there was extensive trade and commerce (much of it in freedman hands) in so large a city, which was also the centre of imperial administration. There was little industry, however, and the urban poor had difficulty finding steady employment. Theirs was a precarious existence, dependent on the public grain dole and on the private charity of the wealthy. Large building programs gave Flavian and Antonine emperors the opportunity not only to repair the damage caused by fire and falling buildings (as stated, a frequent hazard among the densely packed and flimsily built accommodations for the urban plebs) but also to relieve widespread urban unemployment. They also made imperial Rome a city of grandeur. Augustus' building program had been vast but mostly concerned with repairing or rebuilding structures already existing, and his Julio-Claudian successors had built relatively little until the great fire made room for the megalomaniac marvels of Nero's last years. It was under the Flavians and Antonines that Rome obtained many of its most celebrated structures: the Colosseum, Palatine palaces, Trajan's Forum, the Pantheon, the Castel Sant' Angelo (Hadrian's mausoleum), the Temple of Antoninus and Faustina, Aurelius' Column, as well as the aqueducts whose arches spanned across Campagna to keep the city and its innumerable fountains supplied with water.

Italy was much less cosmopolitan and sophisticated and, according to literary tradition, much more sober and straitlaced than was Rome. It was the mistress of the empire, although the gap between it and the provinces was narrowing. Hadrian's policies especially helped to reduce its privileged position. His use of circuit judges was resented precisely because with them Italy resembled a province; actually, Italy badly needed them, and their abolition by Antoninus Pius was soon reversed by Marcus Aurelius. Also, in Aurelius' reign a provincial fate overtook Italy in the form of barbarian invasion; a few years later the country got its first legionary garrison under Septimius Severus.

The economic importance of Italy also declined. By the end of Augustus' reign, the ascendancy of its wine, oil, marble, and fine pottery in the markets of Gaul and Germany had already begun to yield to the competition of local production in the West; and, by Flavian times, Italy was actually importing heavily not only from Gaul (witness the crates of yet-unpacked Gallic bowls and plates caught in the destruction of Pompeii) but also from Spain. The latter province was especially represented by its extraordinarily popular condiment, *garum;* its olive oil, too,

was a sizable item on Italian tables after AD 100, only to yield its primacy there, by the mid-2nd century, to oil from northern Africa. By then, Spanish, Gallic, and African farm products all outweighed Italian ones in Ostia and Rome. Against such tendencies, the emperors did what they could: Domitian, for example, protected Italian viticulture by restricting vine growing in the provinces; Trajan and his successors forced Roman senators to take an interest in the country, even though it was no longer the homeland of many of them, by investing a high proportion of their capital in Italian land (one-third under Trajan, one-quarter under Aurelius).

Attempts to achieve stability in the provinces

Developments in the provinces. The 18th-century historian Edward Gibbon's famous description of the 2nd century as the period when men were happiest and most prosperous is not entirely false. Certainly, by then people had come to take for granted the unique greatness and invincibility of the empire; even the ominous events of Aurelius' reign failed to shatter their conviction that the empire was impregnable; and the internal disturbances of the preceding reign had not given cause for much alarm. The credit for the empire's success lay less with what its rulers did and could do than with what they did not do: they did not interfere too much. The empire was a vast congeries of peoples and races with differing religions, customs, and languages, and the emperors were content to let them live their own lives. Imperial policy favoured a veneer of common culture transcending ethnic differences, but there was no deliberate denationalization. Ambitious men striving for a career naturally found it helpful, if not necessary, to become Roman in bearing and conduct and perhaps even in language as well (although speakers of Greek often rose to exalted positions). But local self-government was the general rule, and neither Latin nor Roman ways were imposed on the communities composing the empire. The official attitude to religion illustrates this—in line with the absolutist trend, emperor worship was becoming slowly but progressively more theocratic (Domitian relished the title of god, Commodus demanded it); yet this did not lead to the suppression of non-Roman or even outlandish cults, unless they were thought immoral (like Druidism, with its human sacrifice) or conducive to public disorder (like Christianity, with its uncompromising dismissal of all gods other than its own as mere demons, and wicked and hurtful ones at that—hence its liability to become a target for riots).

While there is no indication that the central authorities consciously opposed the increase of governmental personnel, the number of government employees certainly grew very slowly. Thus the responsibilities of the magnates in provincial cities were correspondingly great. In parts of southern Spain or in the area south of the Black Sea, for example, where the extent of the territories dependent on cities stretched out over many scores of miles into the surrounding landscape, city senators had not only to collect taxes but also to build roads and carry out much rural police work. Within their cities, too, senators had to see to the collection of taxes and tolls; as a group, they had to oversee and assign the income from municipal lands or buildings rented out and from endowments established by generous citizens; they had to authorize the plans and financing of sometimes very elaborate civic structures—an aqueduct, an amphitheatre, or a temple to the imperial family—or of great annual festivals and fairs or of ongoing amenities serving the public baths (free oil for anointing oneself, heating, and upkeep) or the public markets. In the eastern provinces, they had to replenish from time to time the stock of small local bronze coins; and they had to insure that magistracies were effectively staffed, even though there usually was no salary of any sort to attract candidates. Magistrates and city senators generally had to pay handsomely for their election and thereafter make further handsome contributions, as need arose and so far as they could afford, toward the adornment of their community.

What attracted candidates in adequate numbers were most often three inducements: the feeling of community approval and praise, offered in the most public ways (described by writers of the time with striking psychological penetration); the enhancement of personal influence (meaning power) through the demonstration of great financial means; and finally, the social and political advancement that might follow on local prominence through attracting the attention of a governor or of the emperor himself. It was from the provincial elite that new Roman senators were made.

Cities, through their elite families, competed with each other across entire regions. City rivalries in northern Italy or western Anatolia happen to be especially well reported. Within individual cities, elite families were often in competition as well. In consequence, the standards of municipal beneficence rose, encouraged by a populace who on public occasions assembled in large numbers in the theatre, demanding yet more expenditure from their leaders. The emperors, who realized that the well-being of cities, the jewels of their realm, depended on such munificence, increasingly intervened to insure a continued flow of good things from the rich of a community to their fellow citizens. Legislation might, for example, specify the binding nature of electoral campaign promises or of formerly voluntary contributions connected with public service. As a consequence, in the 2nd century consideration must for the first time be given to the local aristocrat unwilling to serve his city; the series of imperial pronouncements exerting compulsion on such a person to serve was to stretch far into the future, with increasing severity. Attempts to stabilize the benefits arising from ambitious rivalries thus had an oppressive aspect.

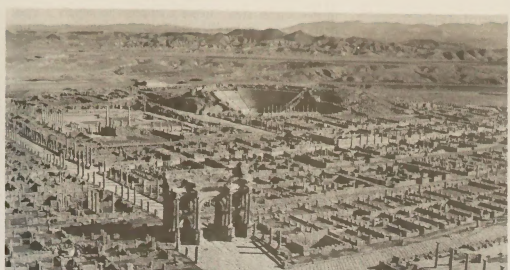
As to the lower orders, their voice is rarely heard in surviving sources, except in acclamation. So long as the rich voluntarily covered the bulk of local expenses and so long as they commanded the leisure and knowledge of the world to give to administration unsalaried, the poor could not fairly claim much of a right to determine the city's choices. Thus they acclaimed the candidacies of the rich and their gifts and otherwise gave vent to their wishes only by shouting in unison in the theatre or amphitheatre (in between spectacles) or through violent mob actions.

As noted above, the poor routinely solved the problems of daily life by appealing to someone of influence locally; this was true whether in Palestine, as indicated in the Talmud, or in Italy, as is evident from Pliny's correspondence. The higher one looked in society, the more it appeared crisscrossed and interconnected by ties of kinship or of past services exchanged. It was at these higher levels that answers to routine problems were to be sought. Appeal was not directed to one's peers, even though trade associations, cult groups of social equals, and burial insurance clubs with monthly meetings could be found in every town. Such groups served social, not political or economic, purposes, at least during the principate.

The social classes

Accordingly, society was ordinarily described by contemporaries simply in terms of two classes, the upper and the lower, rich and poor, powerful and dependent, well known and nameless. The upper classes consisted of little more than 600 Roman senators, 25,000 equites, and 100,000 city senators; hence, a total amounting to 2 percent of the population. This stratum, from the mid-2nd century defined in law as "the more honourable," *honestiores,* was minutely subdivided into degrees of dignity, the degrees being well advertised and jealously asserted; the entire stratum, however, was entitled to receive specially tender treatment in the courts. The remaining population was lumped together as "the more lowly," *humiliores,* subject to torture when giving witness in court; to beatings, not fines; and to execution (in increasingly savage forms of death) rather than exile for the most serious crimes. Yet because of the existing patterns of power, which directed the *humiliores* to turn for help to the upper stratum, the lower classes did not form a revolutionary mass but constituted a stable element.

The pyramidal structure of society suggested by the statistics given above is somewhat obscured by the reality and prominence of the urban scene. In the cities the harsh outlines of the distribution of wealth were moderated by a certain degree of social mobility. No class offers more success stories than that of freedmen. Especially in the West, freedmen are astonishingly prominent in the record



The ancient Roman city of Thamugadi in northeastern Algeria, founded by Trajan in AD 100.
Fototeca Unione

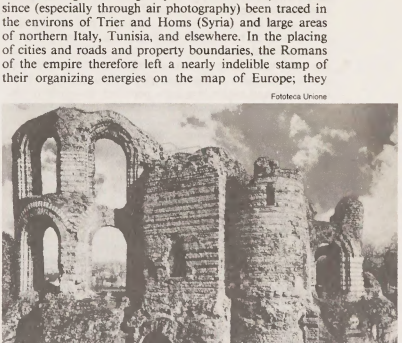
of inscriptions and proverbial for what the upper classes called unprincipled enterprise and vulgar moneygrubbing. Artisans and tradespeople—lowly folk, in the eyes of someone like Cicero—in fact presented themselves with a certain dignity, even some financial ease. At the bottom, slaves were numerous, constituting perhaps one-tenth of the population in at least the larger towns outside of Italy and considerably more in Italy—as much as one-quarter in Rome. But in the cities many of them at least enjoyed security from starvation and had a good roof over their heads. When one turns to the rural scene, however, one encounters a far larger, harsher world. In the first place, nine-tenths of the empire's people lived on the land and from its yield. Where details of their lives emerge with any clarity, they most often tell of a changeless and bleak existence. The city looked down on the countryside with elaborate scorn, keeping the rural population at arm's length. Very often people in the country had their own language—such as Gallic, Syriac, Libyphoenician, or Coptic, which further isolated them—and their own religion, marriage customs, and forms of entertainment. In time, the very term "country dweller," *paganus*, set the rural population still further apart from the empire's Christianized urban population.

In the overall context of Western history, the degree to which the Mediterranean world during the period of the empire became one single system, one civilization, is a matter of the greatest importance. Clearly, one must distinguish between the life of the rural masses and that of the urban minority. The former retained many traits of a way of life predating not only Roman conquest but, in the East, the conquests of Alexander the Great centuries earlier. However, the device of organizing conquered territories under cities responsible for their surrounding territory proved as successful under the Romans as under the Greeks. The intent of both conquerors may have been limited to ensuring political control and the yield of tribute; however, in fact, they achieved much more: an approach to uniformity, at least in the cities.

The first thing to strike the traveler's eye, in any survey of the 2nd-century empire, would have been the physical appearance of urban centres; as already noted, whatever the province, many of the same architectural forms could be observed: the suburbs tended to have aqueducts and racetracks and the cities a central grand market area surrounded by porticoes, temples, a records office, a council hall, a basilica for judicial hearings and public auctions, and a covered market hall of a characteristic shape for perishable foods (a macellum, as in Pompeii, in Perge on the southern coast of modern Turkey, or in North African Lepcis); there also would have been public baths with several separate halls for cold or hot bathing or exercise, a covered or open-air theatre, grand fountains, monumental arches, and honorific statues of local worthies by the dozens or even hundreds. Eastern centres would have gymnasia, occasionally Western ones as well; and Western cities would have amphitheatres, occasionally Eastern ones as well, for the imported institution of gladiatorial combats. Throughout the Western provinces, public buildings were likely to be arranged according to a single plan—more or less the same everywhere—in which a grid of right-angle streets was dominant, at least toward the central part of the city.

In the West, as opposed to the East, a great deal of urbanization remained to be done and was accomplished by the Romans. The grid plan, its particular mark, can be detected at the heart of places such as Turin, Banasa (Morocco), and Autun, all Augustan foundations, as well as in Nicopolis (Bulgaria), Budapest, and Silchester, all later ones. As noted above, orthogonal town planning was not a Roman invention, but the Romans introduced it to new regions and with a particular regularity of their own. Moreover, the grid of the central part of the city was matched, and sometimes extended on the same lines, by another grid laid across the surrounding territory. The process, referred to as centuriation, typically made use of squares of 2,330 feet (710 metres) on a side, intended for land distribution to settlers and general purposes of inventory. Signs of it were first detected in northern Africa in the 1830s, through surviving crop marks and roads, and have since (especially through air photography) been traced in the environs of Trier and Homs (Syria) and large areas of northern Italy, Tunisia, and elsewhere. In the placing of cities and roads and property boundaries, the Romans of the empire therefore left a nearly indelible stamp of their organizing energies on the map of Europe; they

Fototeca Unione



The hot room of the imperial baths at Trier, Ger.

also established the lives of conquered populations inside their own characteristic framework.

The special burst of energy in the Augustan colonizing spread abroad not only the visible elements of a ruling civilization but the invisible ones as well. Colonies and municipalities received Roman forms of government according to their charters, they were administered by Roman law in Latin, and they diffused these things throughout the general population within and around them. In frontier areas such lessons in an alien civilization were pressed home by garrison forces through their frequent contacts with their hosts and suppliers. By the 2nd century considerable Latinization had occurred in the West. Modern Spanish, Portuguese, and French show that this was particularly true of the Iberian peninsula, which had been provincial soil ever since the Second Punic War, and of Gaul, where Latin enjoyed the advantage of some relationship to Celtic. In these regions, except in the less accessible rural or mountainous parts, even the lower orders adopted Latin. Today one can find in Romania the tongue that is the closest to its parent, Latin, even at so great a distance from its home. And Latin can be found not only in Romance languages; it has left its mark on languages such as Basque and German.

Inscriptions represent the most frequent testimony to linguistic allegiance; more than a quarter of a million survive in Latin from the period of the empire, the vast majority of them being funerary. The number of inscriptions per year increases slowly during the 1st century and a half AD, thereafter ascending in a steep line to a point in the second decade of the 3rd and then falling off even more steeply. The curve is best explained as reflecting pride in "Romanness"—in possessing not only Latin but full citizenship as well and, thereby, admission to a group for whom commemoration of the deceased was a legal as well as a moral duty. Over the course of time, by individual gift from the emperors, by army service, and by election to magistracies or simply to the city senates of colonies and municipalities, a growing proportion of the empire's population had gained citizenship; moreover, their children were citizens, whose descendants in turn were Romans in the legal sense. By AD 212 this accelerating process had advanced so far that the emperor Caracalla could offer the gift of incorporation to the entirety of his subjects without much notice being taken of his generosity—it was already in the possession of most of the people who counted and whose reactions might be recorded. Once citizenship was universal, it ceased to constitute a distinction; thus the declaration of it through the custom of funerary commemoration rapidly passed out of favour.

One great flaw in the picture of the empire as one single civilization by 212, triumphantly unified in culture as in its political form, has already been pointed out: what was achieved within the cities' walls did not extend with any completeness to the rural population, among whom local ways and native languages persisted. Peasants in 4th-century Syria spoke mostly Syriac, in Egypt mostly Coptic, in Africa often Punic or Libyphoenician, and in the Danube and northwestern provinces other native tongues. There was still another great flaw: the empire was half Roman (or Latin), half Greek. The latter was hardly touched by the former except through what may be called official channels—that is, law, coinage, military presence, imperial cult, and the superposition of an alien structure of power and prestige, to which the elite of the Eastern provinces might aspire. On the other hand, the Roman half was steeped in Greek ways. Apuleius, for example, though born and reared in a small North African town of the 2nd century, was sent to Athens to study rhetoric; on his return he could find not only an audience for his presentations in Greek but ordinary people in the marketplace able to read a letter in that language. In Rome the Christian community used Greek as its liturgical language well into the 3rd century, and the crowds in the Circus Maximus could enjoy a pun in Greek; an aristocrat such as the emperor Marcus Aurelius could be expected to be as bilingual as was Cicero or Caesar before him or even, like the emperor Gallienus, help the Greek philosopher Plotinus found a sort of Institute for Advanced Studies in the Naples area. Greece continued to supply a great deal of sculpture for Western buyers or even the teams of artisans needed for the decoration of public buildings in 3rd-century northern Africa. By such various means the division between the two halves of the empire was for a time covered over.

Among the institutions most important in softening the edges of regional differences was the cult of the emperors. In one sense, it originated in the 4th century BC, when Alexander the Great first received veneration by titles and symbols and forms of address as if he were a superhuman being. Indeed, he must have seemed exactly that to contemporaries in Egypt, where the pharaohs had long been worshiped, and to peoples in the Middle East, for similar reasons of religious custom. Even the Greeks were quite used to the idea that beings who lived a human life of extraordinary accomplishment, as "heroes" in the full sense of the Greek word, would never die but be raised into some higher world; they believed this of heroes such as Achilles, Hercules, Pythagoras, and Dion of Syracuse in the mid-4th century BC. Great Roman commanders, like Hellenistic rulers, had altars, festivals, and special honours voted to them by Greek cities from the start of the 2nd century BC. It was not so strange, then, that a freedman supporter of Caesar's erected a pillar over the ashes of the dead dictator in the Forum in April 44 BC and offered cult to him as a being now resident among the gods. Many citizens joined in. Within days Caesar's heir Octavian pressed for the declaration of Caesar as divine—which the Senate granted by its vote in 42. By 25 BC the city of Mytilene had organized annual cult acts honouring Augustus and communicated their forms and impulse to Tarraco in Spain as well as to other Eastern Greek cities; and by 12 BC divine honours to Caesar and Augustus' *genius* were established through the emperors' initiative both in the Gallic capital, Lugdunum, and in the neighbourhood chapels to the crossroads gods in Rome. From these various points and models, emperor worship spread rapidly. Within a few generations, cities everywhere had built in its service new temples that dominated their forums or had assigned old temples to the joint service of a prior god and the imperial family. Such centres served as rallying points for the citizenry to express its devotion to Rome and the emperor. To speak for whole provinces, priests of the cult assembled during their year of office in central shrines, such as Lugdunum, as delegates of their cities, where they formulated for the emperor their complaints or their views on the incumbent governor's administration. Whether these priests were freedmen in urban neighbourhoods, municipal magnates in local temples, or still grander leaders of the provinces, they perceived the imperial cult as something of high prestige and invested it and Roman rule with glory.

The cult of the emperors

The emotional and political unification of the empire was further promoted by submissive or flattering forms of reference or address, adopted even by the highest personages when speaking of the emperor, and by portraits of the emperors or their families with attendant written messages. Of these two most obvious means of propaganda, the first survives in the texts of many panegyrics delivered to the throne, rhetorical disquisitions on monarchy, and prefatory announcements accompanying the publication of government edicts. They established a tone in which it was proper to think of Roman rule and government. Portraits, the second means of propaganda, included painted ones on general display in cities, sculpted ones, especially in the early years of each reign, based on official models available in a few major cities (hundreds of these survive, including at least one in gold), and engraved ones on coins. Imperial coins offered a more rapidly changing exhibition of images than even postage stamps in the modern world. Because the dies soon wore out, many scores of issues had to be brought out each year, in gold, silver, and bronze. While the images ("types") and words ("legends") on them tended to repetition, there was much conscious inculcation of topical messages: for example, in the short and rocky reign of Galba in AD 69, one finds the legends "All's well that ends well" (*bonus eventus*), "Rome reborn," "Peace for Romans," and "Constitutional government restored"

Portraits of the emperor

(*libertas restituta*, with iconographic reference to Brutus' coins of 43 BC) and superlative portraits of Galba himself; or, in other reigns, the legends, enriched with suitable symbolism, read "the soldiers loyal," "Italy well fed," and *fecunditas* of the royal family and its progeny. So far as it is possible to comprehend the mind of the empire's populace, there was no significant opposition to the government by the 2nd century; instead, there prevailed a great deal of ready veneration for the principate as an institution.

Economic factors, to the extent that they were favourable, played an obvious part in promoting both cultural and political unity. So far as acculturation was concerned, a limit to its achievement was clearly set by the amount of disposable capital among non-Romanized populations. The cost of such luxuries as schooling in Latin or frescoes on one's walls were high. But more and more people could afford them as the benefits of Roman occupation were spreading. The rising levels of prosperity did not, however, result from a special benevolence on the part of the conquerors, intent as they were (and often cruelly intent) on the pleasures and profits of physical mastery over the conquered. Rather, they can be explained, first, by the imposition of the Pax Romana, which gave urban centres surer access to the surrounding rural areas and rural producers access in turn to convenient, centralized markets; second, by the sheer attractiveness of imported articles, which intensified efforts to increase the power to buy them; third, by the economic stimulation afforded by taxes, which had to be paid on new earnings but which remained in the provinces where they were raised. In the fourth place, prosperity also rose in the regions least Romanized. This can be explained by the fact that they tended to be heavily garrisoned and the soldiers spent their wages locally. So far as they could, they bought goods and services of a Roman sort and generally attracted concentrations of people likely to develop into cities of a Roman sort. The economic impact of army payrolls was all the greater because of the cash added to them from taxes raised in other, more developed provinces in the East. Much of the urbanization and enrichment of the western and northern provinces can be explained by these four factors.

Until the 1950s or '60s the sources for studying the economy of the empire were insufficient. The archaeological sources were too scarce and heterogeneous to be of much help, and the written ones contained barely usable amounts of quantified data; economic analysis without quantification, however, is almost a contradiction in terms. Thus discussion was obliged to limit itself to rather general remarks about the obviously wide exchange of goods, the most famous points of production or sale of given articles, techniques of banking, or commercial law. This is still the case with regard to the Eastern half of the Mediterranean world, where excavation has made relatively little headway; but, for the West, archaeological data have greatly increased in recent decades in both quantity and intelligibility. As a result, a growing number of significant statements based on quantification can now be made. They are of special value because they bear on what was economically most important—namely, agriculture. Like any preindustrial economy, that of the empire derived the overwhelming bulk of its gross national product from food production. One would therefore like to know what regions in what periods produced what rough percentage of the chief comestibles—wine, oil, wheat, *garum* or legumes. Thanks to techniques such as neutron activation analysis or X-ray fluorescence spectrometry, the contents of large samples of amphorae at certain market junctures can be identified, dated by shape of vessel, and occasionally ascribed to certain named producers of the vessel, and the information drawn into a graph; or, the numbers and find-spots of datable fine "china" (so-called Arretine ware or later equivalents) or ceramic oil lamps from named producers can be indicated on a map of, say, Spain or France. The yield of such data underlies statements made above regarding, for example, the supersession of Italy as producer of several essential agricultural products by the mid-1st century AD, the concurrent transformation of Gaul from importer to exporter, and the emergence by the 3rd century of northern Africa as a major exporter of certain very common articles. Information of this general nature provides some sense of the shift in prosperity in the Western provinces.

Archaeological sources for studying the economy

In the age of the Antonines, Rome's empire enjoyed an obvious and prosperous tranquility; modern consensus has even settled on about AD 160 as the peak of Roman civilization. Whatever measurement may be used in this identification, however, an economic one does not fit very well. Evidence, as it accumulates in more quantifiable form, does not seem to show any perceptible economic decline in the empire as a whole after roughly 160. Rather, Italy had probably suffered some decrease in disposable wealth in the earlier 1st century. Gaul's greatest city, Lugdunum, had begun to shrink toward the end of the 2nd, and various other regions in the West suffered setbacks at various times, while all of Greece continued to be poor. Other regions, however, had more wealth to spend, and as is manifest in major urban projects of utility and beautification or in the larger rooms and increasingly expensive decoration of rural villas. Roman rule also brought extraordinary benefits to the economies of Numidia and Britain, to name its two most obvious successes.

To the extent the empire grew richer, modern observers are likely to look for an explanation in technology. As noted above, in Augustus' reign a new mode of glass-blowing spread rapidly from Syria to other production centres; Syria in the 3rd century was also the home of new and more complicated weave patterns. Such rather minor items, however, only show that technical improvements in industry were few and insignificant. The screw press for wine and olive oil was more efficient than the levered variety, but it was not widely adopted, even within Italy. Waterwheels for power, known in Anatolia in Augustus' reign, were little used; a few examples in Gaul belong only to the later empire. Similarly, the mechanical reaper was found only in Gaul of the 4th century. Perhaps the most significant advances were registered in the selective breeding of strains of grains and domestic animals: for example, the "Roman" sheep (which had originated in the Greek East) spread throughout Europe, banishing the inferior Iron Age species to a merited exile in the Outer Hebrides (the Soay sheep of St. Kilda island). What is vastly more significant, however, than these oddments of technological history is the minute subdivision of productive skills and their transmission from father to son in populations adequate to the demand—for iron ore from Noricum, most notably, or for glass and paper from Alexandria. Specialization in inherited skills produced a remarkably high level of proficiency, requiring only the security of the Pax Romana for the spreading of its products everywhere—transport itself being one of those skills.

The health of the economy no doubt helps to explain the political success of the empire, which was not disturbed by frequent revolts or endemic rural or urban unrest. On the other hand, there were limits in the economy, which expressed themselves through resistance to taxation. Tax levels settled at the enforceable maximum; but revenue fell far short of what one might expect, given the best estimates of the empire's gross national product. The basic problem was the tiny size of the imperial government and the resulting inefficiency of its processes. Moreover, it could not make good its inadequacies by borrowing in times of special need: Nero's need to harry his millionaire subjects with false charges of treason in order to pay for his incredibly expensive court and spendthrift impulses reflects the realities of raising revenue. So do the very cautious experiments of Augustus in setting army pay and army size. Ultimately, the military strength of the empire was insufficient—inadequate for emergencies—because of these realities.

The army. The army that enforced the Pax Romana had expanded little beyond the size envisaged for it by Augustus, despite the enlargement of the empire by Claudius, the Flavians, and Trajan. It reached 31 legions momentarily under Trajan, but it usually numbered 28 under the Flavians and Antonines until the onset of the frontier crisis in Aurelius' reign brought it to 30. Without raising pay rates to attract recruits more easily, a large force was

seemingly beyond reach—which probably explains why Hadrian, and later Commodus, halted further expansion.

The army was used not to prop up a militarist government but to defend the frontiers. Shifts in enemy pressures, however, caused the legions to be distributed differently than in Julio-Claudian times. Under Antoninus Pius, the Danubian provinces (Pannonia, Moesia, Dacia) had 10, and the East (Anatolia, Syria, Palestine, Egypt) had 9, and both regions also had supporting naval flotillas; of the remaining 9 legions, Britain contained 3 and the Rhineland 4. Tacitus in his *Annals* (4.5) rates the auxiliary troops near the turn of the era as being about as numerous as the legionaries. But they soon outnumbered them: that is, whereas legions contained somewhat more than 5,000 men each if they were at full strength and thus totaled roughly 150,000 in the mid-2nd century, the auxiliaries numbered 245,000—again, if at full strength. Recent estimates put the actual figure for the entire army at 375,000 to 400,000.

Two reasons, military and financial, explain the growing use of nonlegionaries. Mustered in units mostly of 500, they were easier to move around and could be encouraged to maintain the special native skills of their inheritance—as slingers from the Balearic Islands or Crete, in camel corps from Numidia, or as light cavalry from Thrace. In addition, they could be recruited for lower wages than legionaries. As regards recruitment for the legions, even that higher rate proved less and less attractive. Whereas legions in the early empire could be largely filled with men born in Italy and southern Gaul, by the second half of the 1st century most of the men had to be drawn from the provinces; after Trajan, they were largely natives of the frontier provinces. Young men from the inner parts of the empire, growing up in successive generations of continual peace, no longer looked on military service as a natural part of manhood, and the civilian economy appeared attractive compared to the rewards at some frontier posting. Peace and prosperity thus combined to make the army less and less Roman, less and less of the centre, and more and more nearly barbarous.

The troops' loyalty did not suffer on that account. The men were no more ready to mutiny or to support a pretender around AD 200 than they had been in the early empire. However, experience especially in the year of the four emperors (AD 69) did suggest the desirability of splitting commands into smaller units, which, in turn, involved splitting up provinces, the number of which was constantly growing: by Hadrian's day subdivision began to anticipate the fragmentation later carried out by Diocletian.

Imperial literature

Cultural life. The literature of the empire is both abundant and competent, for which the emperors' encouragement and financing of libraries and higher education were perhaps in part responsible. The writers, however, with the possible exception of Christian apologists, were seldom excitingly original and creative. As Tacitus said, the great masters of literature had ceased to be. Perhaps Augustus' emphasis on tradition affected more than political ideals and practice. At any rate, men of letters, too, looked often backward. At the same time, they clearly reveal the success of the empire in spreading Greco-Roman culture, for the majority of them were natives of neither Italy nor Greece. Of the writers in Latin, the two Senecas, Lucan, Martial, Columella, Hyginus, and Pomponius Mela came from Spain; Fronto, Apuleius, and probably Florus and Aulus Gellius, from Africa. Tacitus was perhaps from Gallia Narbonensis. The Latin writers in general sought their models less in Greece than in Augustus' Golden Age, when Latin literature had reached maturity. Thus, the poets admired Virgil and imitated Ovid; lacking genuine inspiration, they substituted for it an erudite cleverness, the fruit of an education that stressed oratory of a striking but sterile kind. Authentic eloquence in Latin came to an end when, as Tacitus put it, the principate "pacified" oratory. Under the Flavians and Antonines, an artificial rhetoric, constantly straining after meretricious effects, replaced it. The epigrammatic aphorism (*sententia*) was especially cultivated; the epics of Lucan, Valerius Flaccus, Silius Italicus, and Statius are full of it, and it found a natural outlet in satirical writing, of which the Latin instinct for the mordant always ensured an abundance. In fact, Latin satire excelled: witness Martial's epigrams, Petronius' and Juvenal's pictures of the period, and Persius' more academic talent. For that matter, Tacitus' irony and pessimism were not far removed from satire.

In the East the official status of Greek and the favour it enjoyed from such emperors as Hadrian gave new life to Greek literature. It had something in common with its Latin counterpart in that it looked to the past but was chiefly written by authors who were not native to the birthplace of the language. The so-called Second Sophistic reverted to the atticism of an earlier day but often in a Roman spirit; its products from the Asian pens of Dio Chrysostom and Aelius Aristides are sometimes limpid and talented tours de force but rarely great literature. In Greek, too, the best work was in satire, the comic prose dialogues of the Syrian Lucian being the most noteworthy and original literary creations of the period. Among minor writers the charm of Arrian and Pausanias, Asians both, and above all of Plutarch abides (although Plutarch's talents were mediocre, and his moralizing was shallow, his biographies, like those of his Latin contemporary Suetonius, are full of information and interest).

Imperial encouragement of Greek culture and a conviction, no longer justified, of its artistic and intellectual superiority caused the East to resist Latinization. This attitude was bound to lead to a divided empire, and thoughtful observers must have noted it with misgivings. The split, however, was still far in the future. Meanwhile, there was a more immediate cause for disquiet. The plethora of summaries and anthologies that appeared implies a public progressively indifferent to reading whole works of literature for themselves. In other words, the outlook for letters was poor, and this had an unfortunate effect on the scientific literature of the age, which was in itself of first-class quality. Dioscorides on botany, Galen on medicine, and Ptolemy on mathematics, astronomy, and geography represent expert scholars expounding carefully, systematically, and lucidly the existing knowledge in their respective fields. But their very excellence proved fatal because, as the reading public dwindled, theirs remained standard works for far too long; their inevitable errors became enshrined, and their works acted as brakes on further progress.

Stoicism was the most flourishing philosophy of the age. In the East a sterile scholasticism diligently studied Plato and Aristotle, but Epictetus, the stoic from Anatolia, was the preeminent philosopher. In the West, stoicism permeates Seneca's work and much of Pliny's *Natural History.* Evidently, its advocacy of common morality appealed to the traditional Roman sense of decorum and duty, and its doctrine of a world directed by an all-embracing providence struck a responsive chord in the 2nd-century emperors, though they deeply disapproved of its extremist offshoots, the cynics: Marcus Aurelius, as noted, was himself a stoic.

Imperial art

Imperial art, dealing above all with man and his achievements, excelled in portraits and commemoration of events; Roman sculpture and presumably Roman painting, also, owed much to Greek styles and techniques. It emerged, however, as its own distinctive type. The Augustan age had pointed the way that Roman art would go: Italian taste would be imposed on Hellenic models to produce something original. The reliefs of the Augustan Ara Pacis belong to Rome and Italy, no matter who actually carved them. By Flavian times this Roman artistic instinct had asserted itself and with it the old Roman tendency toward lively and accurate pictorial representation. It can be seen from the reliefs illustrating the triumph over Judaea in the passageway of the Arch of Titus in the Roman Forum. The narrative description dear to Roman art found its best expression in the great spiral frieze on Trajan's Column, where the emperor can be seen among his soldiers at various times in the Dacian campaigns; the story of the war plays a most important part, although, like most imperial monuments, the column is meant to exalt the leader. Under Hadrian a reaction made sculpture less markedly Italian, as if to be in conformity with the slow decline of Italy toward quasi-provincial status. Also under Hadrian, the figure of the emperor was more prominent—bigger

and more frontal than the other figures—as if to illustrate the growing monarchical tone of the principate. This tendency continued under the Antonines, when there was a magnificent flowering of sculpture on panels, columns, and sarcophagi; but its exuberance and splendour foreshadow the end of classical art.

The artistic currents that flowed in Rome were felt throughout the empire, the less developed areas being influenced most. In the West, provincial sculpture closely resembled Roman, although it sometimes showed variations, in Gaul especially, owing to local influences (the native element, however, is not always easy to identify). The Roman quality of portraits painted on Egyptian mummy cases shows that the Greek-speaking regions were also affected, although generally they maintained their own traditions. But by now the Greek East had become rather barren; much of its production was imitative rather than vitally creative. Greece proper contributed little, the centre of Hellenism having shifted to Anatolia, to places such as Aphrodisias, where there was a flourishing school of sculpture.

In at least one respect the East was heavily influenced by Rome. The use of concrete and cross vault enabled Roman architects and engineers to span wide areas; their technological achievements included the covered vastness of the huge thermal establishments, the massive solidity of the amphitheatres, and the audacity of the soaring bridges and aqueducts. The East was greatly impressed. Admittedly, the agoras and gymnasiums in Greek towns are hardly Roman in aspect, but, for most structures of a practical utilitarian kind, the Greek debt to Rome was heavy. Sometimes Roman influence can be seen not only in the fundamental engineering of such buildings as market gateways, theatres, and amphitheatres but even in such decorative details as composite capitals as well. Roman features abound in exotic Petra, Palmyra, Gerasa, and Baalbek, and even in Athens itself. (E.T.S./R.MacM.)

THE LATER ROMAN EMPIRE

The dynasty of the Severi (AD 193–235). *Septimius Severus.* After the assassination of Commodus on Dec. 31, AD 192, Helvius Pertinax, the prefect of the city, became emperor. In spite of his modest birth, he was well respected by the Senate, but he was without his own army. He was killed by the praetorians at the end of March 193, after a three-month reign. The praetorians, after much corrupt bargaining, designated as emperor an old general, Didius Julianus, who had promised them the largest *donativum* (a donation given to each soldier on the emperor's accession). The action of the praetorians roused the ire of the provincial armies. The army of the Danube, which was the most powerful as well as the closest to Rome, appointed Septimius Severus in May 193. Severus soon had to face two competitors, supported, like himself, by their own troops: Pescennius Niger, the legate of Syria, and Clodius Albinus, legate of Britain. After having temporarily neutralized Albinus by accepting him as Caesar (heir apparent), Septimius marched against Niger, whose troops, having come from Egypt and Syria, were already occupying Byzantium. The Danubian legions were victorious, and Niger was killed at the end of 194; Antioch and Byzantium were pillaged after a long siege. Septimius even invaded Mesopotamia, for the Parthians had supported Niger. But this campaign was quickly interrupted: in the West, Albinus, disappointed at not being associated with the empire, proclaimed himself Augustus in 196 and invaded Gaul. He was supported by the troops, by the population, and even by the senators in Rome. In February 197 he was defeated and killed in a difficult battle near his capital of Lugdunum, which, in turn, was almost devastated. Septimius Severus remained the sole master of the empire, but the pillagings, executions, and confiscations left a painful memory. A few months later, in the summer of 197, he launched a second Mesopotamian campaign, this time against the Parthian king Vologases IV, who had attacked the frontier outpost Nisibis conquered two years previously by the Romans. Septimius Severus was again victorious. Having arrived at the Parthian capitals (Seleucia and Ctesiphon), he was defeated near Hatra but in 198 obtained an advantageous peace: Rome retained a part of Mesopotamia, together with Nisibis, the new province being governed by an eques. After having inspected the East, the emperor returned to Rome in 202. He spent most of his time there until 208, when the incursions of Caledonian rebels called him to Britain, where he carried out a three-year campaign along Hadrian's Wall. He died at Eboracum (York) in February 211.

Severus' campaigns

Septimius Severus belonged to a Romanized Tripolitan family that had only recently attained honours. He was born in Leptis Magna in North Africa and favoured his native land throughout his reign. He was married to Julia Domna of Emesa, a Syrian woman from an important priestly family, and was surrounded by Easterners. He had pursued a senatorial career and had proved himself a competent general, but he was above all a good administrator and a jurist. Disliking Romans, Italians, and senators, he deliberately relied on the faithful Danubian army that had brought him to power, and he always showed great concern for the provincials and the lower classes. Although he had sought to appropriate the popularity of the Antonines to his own advantage by proclaiming himself the son of Marcus Aurelius and by naming his own son Marcus Aurelius Antoninus, he in fact carried out a totally different policy—a brutal yet realistic policy that opened careers to new social classes. Indifferent to the prestige of the Senate, where he had a great many enemies, he favoured the equites. The army thus became the seedbed of the equestrian order and was the object of all of his attentions. The ready forces were increased by the creation of three new legions commanded by equites, and one of these, the Second Parthica, was installed near Rome. Unlike Vespasian, who also owed his power to the army but who knew how to keep it in its proper place, Septimius Severus, aware of the urgency of external problems, established a sort of military monarchy. The praetorian cohorts doubled their ranks, and the dismissal of the old staff of Italian origin transformed the Praetorian Guard into an imperial guard, in which the elite of the Danube army were the most important element. The auxiliary troops were increased by the creation of 1,000-man units (infantry cohorts) and cavalry troops, sometimes outfitted with mail armour in the Parthian manner. The careers of noncommissioned officers emerging from the ranks now opened onto new horizons: centurions and noncommissioned grades could attain the tribunate and enter into the equestrian order. Thus, a simple Illyrian peasant might attain high posts: this was undoubtedly the most significant aspect of the "Severan revolution." This "democratization" was not necessarily a barbarization, for the provincial legions had long been Romanized. Their salaries were increased, and *donativa* were distributed more frequently; thenceforth, soldiers were fed at the expense of the provincials. Veterans received lands, mostly in Syria and Africa. The right of legitimate marriage, previously refused by Augustus, was granted to almost all of the soldiers, and the right to form *collegia* (private associations) was given to noncommissioned officers. Because more than a century had passed since the last raise in pay for the troops, despite a steady (if slow) rise in the level of prices, Severus increased the legionary's base rate from 300 to 500 *denarii*, with, no doubt, corresponding increases in other ranks. The reflection of this step in the content of precious metal in silver coinage (see below) recalls a point made earlier: the imperial revenues were constrained within the narrow limits of political and administrative reality.

The growth of the equestrian order

The administrative accomplishments of Septimius Severus were of great importance: he clearly outlined the powers of the city prefect; he entrusted the praetorian prefecture to first-class jurists, such as Papinian; and he increased the number of procurators, who were recruited for financial posts from among Africans and Easterners and for government posts (*praesides*) from among Danubian officers. Italy lost its privileges and found itself subjected, like all the other provinces, to the new *annona*, a tax paid in kind, which assured the maintenance of the army and of the officials. The consequent increase in expenditures—for administration, for the salaries and the *donativa* of the soldiers, for the maintenance of the Roman plebs, and for

Severus' administrative accomplishments

construction—obliged the emperor to devalue the denarius in 194. But the confiscations increased his personal fortune, the *res privata,* which had been previously created by Antoninus.

Severus' social policy favoured both the provincial recruitment of senators (Easterners, Africans, and even Egyptians), causing a sharp decrease in the percentage of Italian senators, and the elevation of the equestrian order, which began to fill the prince's council with its jurists. The cities, which had been favoured by the Antonines, were more and more considered as administrative wheels in the service of the state: the richest *decuriones* (municipal councillors) were financially responsible for levying the taxes, and it was for this purpose that the towns of Egypt finally received a *boulē* (municipal senate).

The burden of taxes and forced government service was made weightier by numerous transport duties for the army and for the *annona* service and was regulated by the jurists through financial, personal, or mixed charges. The state was watchful to keep the *decuriones* in the service of their cities and to provide a control on their administration through the appointment of *curatores rei publicae,* or officials of the central government. The lower classes were, in principle, protected against the abuses of the rich, but in fact they were placed at the service of the state through the restrictions imposed on shipping and commercial corporations. Membership might entail forced contributions of capital or labour to such public necessities as the supply of food to Rome. The state became more and more a policeman, and the excesses of power of numerous grain merchants (*frumentarii*) weighed heavily on the little man.

Imperial power, without repudiating the ideological themes of the principate, rested in fact on the army and sought its legitimacy in heredity: the two sons of Septimius Severus, Caracalla and Geta, were first proclaimed Caesars, the former in 196, the latter in 198; later, they were directly associated with imperial power through bestowal of the title of Augustus, in 198 and 209, respectively. Thus, during the last three years of Septimius Severus' reign, the empire had three Augusti at its head.

Caracalla. Caracalla, the eldest son of Septimius Severus, reigned from 211 to 217, after having assassinated his younger brother, Geta. He was a caricature of his father: violent, megalomaniacal, full of complexes, and, in addition, cruel and debauched. He retained the entourage of the equites and jurists who had governed with his father but enforced to an even greater degree his father's militaristic and egalitarian policy. He increased the wages of the army even further and, at the same time, began a costly building program that quickly depleted the fortune left him by his father. He forced the senators to pay heavy contributions, doubled the inheritance and emancipation taxes, and often required the *aurum coronarium* (a contribution in gold), thereby ruining the urban middle classes. To counter the effects of a general upward drift of prices and the larger and better-paid army of his own and his father's making, he created a new silver coin, the antoninianus. It was intended to replace the basic denarius at double its value, though containing only about one and a half times its worth in precious metal. The only historical source to suggest Caracalla's motive for his gift of universal citizenship, Dio Cassius, states that it was meant to increase revenues by bringing new elements of the population under tax obligations formerly limited to Romans only.

The antoninianus

Although little endowed with military qualities, Caracalla adopted as his patron Alexander the Great, whom he admired greatly, and embarked on an active external policy. He fought successfully against the Teutonic tribes of the upper Danube, among whom the Alamanni, as well as the Capri of the middle Danube, appeared for the first time; he often prudently mixed military operations with negotiation and gave important subsidies and money (in sound currency) to the barbarians, thus arousing much discontent. His ambition was to triumph in the East like his hero of old and, more recently, Trajan and his own father. He invaded Armenia and Adiabene and annexed Osroëne in northwest Mesopotamia, joining it to the part of Mesopotamia taken by Septimius Severus. In April 217, while pursuing his march on the Tigris, he was assassinated on the order of one of his praetorian prefects, Marcus Opellius Macrinus.

Macrinus. Macrinus was accepted as emperor by the soldiers, who were unaware of the role he had played in the death of his predecessor. For the first time an eques had acceded to the empire after having been no more than a manager of financial affairs. The senators reluctantly accepted this member of the equestrian order, who, nevertheless, proved to be moderate and conciliatory; but the armies despised him as a mere civilian, and the ancient authors were hostile to him. His reign was brief, and little is known of him. He concluded an inglorious peace with the Parthians, which assured Mesopotamia to Rome through the payment of large sums of money. And to make himself popular, he canceled Caracalla's tax increases and reduced military expenditures. A plot against him was soon organized: two young grandnephews of Septimius Severus were persuaded by their mothers and especially by their grandmother, Julia Maesa, the sister of Julia Domna (who had recently died), to reach for imperial power. The eldest, Bassianus, was presented to the troops of Syria, who had been bought with gold, and was proclaimed in April 218. Shortly afterward, Macrinus was defeated and killed, as was his son (whom he had associated with him on the throne).

Elagabalus and Severus Alexander. The new emperor was presented as the son of Caracalla, whose name he took (Marcus Aurelius Antoninus). He is better known, however, under the name Elagabalus, the god whose high priest he was and whom he quickly and imprudently attempted to impose on the Romans, in spite of his grandmother's counsel of moderation. Fourteen years old, he caused himself to be detested by his heavy expenditures, his orgies, and the dissolute behaviour of his circle. The praetorians killed him in 222 and proclaimed as emperor his first cousin, Alexianus, who took the name of Severus Alexander.

Although well educated and full of good intentions, Severus Alexander showed some weakness of character by submitting to the counsel of his mother, Mamaea, and of his grandmother, Maesa. The *Scriptores historiae Augustae,* a collection of biographies of the emperors, attributes to him a complete program of reforms favourable to the Senate, but these reforms are not mentioned elsewhere. As in the time of Septimius Severus, his counselors were equites. Ulpian, the praetorian prefect, was the greatest jurist of this period, and the basic policies of the founder of the dynasty were carried on, but with less energy. This weakening of energy had disastrous results: in Persia, the Arsacids were replaced in 224 by the more ambitious Sāsānid dynasty, who hoped to recover the former possessions of the Achaemenids in the East. Their initial attacks were stopped in 232 by a campaign that was, however, poorly conducted by the emperor and that alienated the army as a result of its ineptitude. In Rome there were frequent disorders, and, as early as 223, Ulpian had been killed by the praetorians. While gathered on the Rhine to fight the Teutons, the soldiers once again revolted and killed Severus Alexander and his mother. A coarse and uneducated but energetic soldier, Maximinus the Thracian, succeeded him without difficulty in March 235. The Severan dynasty had come to an end.

Rise of the Sāsānids in Persia

Religious and cultural life in the 3rd century. On the right bank of the Tiber in Rome, in the least fashionable section of town among Lebanese and Jewish labourers, Elagabalus built an elegant temple to his ancestral god; he was no doubt in those precincts very well received when he presided personally at its inauguration. Yet the world that counted, the world of senators and centurions, reacted with indignation. Within the capital the ruler was expected to honour the gods of the capital, the ancient Roman ones. At the same time, it was deemed appropriate that he reverently recognize other gods, in their place; for this reason a biography presenting Severus Alexander for the reader's admiration records how scrupulously he offered worship on the Capitoline to Jupiter, while also having, in a chapel attached to his domestic quarters, the images of his *lares* (household gods), of the deified emperors of most

beloved memory, and of such superhuman beings as the Greeks would have called "heroes," including Apollonius the holy man of Tyana, Christ, Abraham, and Orpheus. The furnishing of the chapel is described by a most dubious source; but if it is not history, it is at least revealing of ideals. A Roman ruler was to express not only the piety of the capital and its citizens but also that of all his people throughout his empire. Imperial religion was properly compounded of both Roman and non-Roman piety.

The issue of an offical religion

Official religion can hardly be said to have existed in the sense of being pressed on people by the state. But the statement needs qualification. The cults of Rome were certainly official in the city itself; they were supported out of the state treasury and by the devotion of the emperor, at least if he lived up to what everyone felt were his responsibilities. In the army, too, camps had shrines in which portraits of the emperor were displayed for veneration on certain days of the year. A 3rd-century calendar has been found in an Eastern city that specifies for the garrison regiment the religious ceremonies to be carried out during the year, including a number of the oldest and most traditional ones in Rome. Many Western cities accorded special size and prominence to a temple in which Jupiter or the imperial family or both together were worshiped not by orders from on high, it is true, but spontaneously. The ubiquity of the imperial cult has already been emphasized. All these manifestations of piety gave some quality of "Romanness" to the religion of the empire.

On the other hand, the empire had been assembled from a great number of parts, whose peoples already had their own way of life fully matured; they were not about to surrender it nor, in fact, were they ever asked to do so by their conquerors. What characterized the religious life of the empire as a whole was the continued vitality of local cults in combination with a generally reverent awareness of one's neighbours' cults. The emperor, for example, might openly offer personal veneration to his favourite god, a god outside the traditional Roman circle, while also practicing a more conventional piety. When he was on his travels, he would offer cult at the chief shrines of all the localities he visited. What was expected of the emperor was expected of everyone: respectful toleration of all components in the religious amalgam. Of course, there were differences according to individual temperament and degree of education; approaches to religion might be literal or philosophical, fervent or relaxed. Rural society was more conservative than urban. But the whole can fairly be called an integrated system.

Just as the special power of the Greek gods had gained recognition among the Etruscans and, subsequently, among the Romans in remote centuries BC or as Sarapis in Hellenistic times had come to be worshiped in scattered parts of the Ptolemies' realm—Macedonia and Ionia, for example—so at last the news of unfamiliar gods was carried by their worshipers to distant places in the Roman Empire where, too, they worked their wonders, attracted reverent attention, and received a pillared lodging, a priesthood, and daily offerings. The Pax Romana encouraged a great deal more than commerce in material objects. It made inevitable the exchange of ideas in a more richly woven and complex fabric than the Mediterranean world had ever seen, in which the Phrygian Cybele was at home also in Gaul and the Italian Silvanus in northern Africa.

Religious developments in the Eastern provinces during the centuries from Augustus to Severus Alexander followed a somewhat different course from those in the West. In the East the further jumbling together of already well-mixed traditions encouraged a tolerance that eroded their edges. It became possible to see predominant similarities in Selene, Artemis, and Isis, in Zeus, Iarhibol, Helios, and Sarapis, or in Cybele, Ma, and Bellona. From recognition of basic similarities one might reason to a sort of monotheism, by the lights of which, for persons given to theology, local deities were no more than narrow expressions of greater truths. A juncture was then natural with Neoplatonism, the school of philosophy that later came to be held in high regard.

On the other hand, in Italy, the Danube provinces, and the Western provinces, religious change and development can be more easily seen in the immigration of worshipers of Easter deities. Those took root and became popular—none more so than Mithra, though Isis, Cybele, and Jupiter of Doliche were close behind. Apuleius in the closing chapters of his novel usually called *The Golden Ass* in English describes how a young man is brought from mere consciousness of Isis as a famous goddess with certain well-known rites and attributes, to a single-minded devotion to her. Aelius Aristides, a famous rhetorician of the time, recounts in his spiritual diary the development of a similar devotion in himself to Asclepius. Both the fictional and the factual account give a central place to benefits miraculously granted. It was by such means that piety was ordinarily warmed to a special fervour, whether or not that process should be called conversion. In any case, it produced the testimonies—votive inscriptions, temples, and so forth—through which it is possible to trace the spread of foreign cults. Eastern cults, however, also introduced to the West complex liturgies, beliefs underlying beliefs that could be explained in especially dramatic ways to special devotees ("mysteries"), and much rich symbolism. Of no cult was this more true than Mithraism, known to the 20th century through excavation of the underground shrines that it preferred.

The rise of Christianity. During the 1st and 2nd centuries, Christianity spread with relative slowness. The doctrines of Jesus, who was crucified about AD 30, first took root among the Jews of Palestine, where a large number of sects were proliferating—orthodox sects, such as the Sadducees and the Pharisees, as well as dissident and sometimes persecuted sects such as the Essenes, whose ascetic practices have been illuminated by the discovery of the Dead Sea Scrolls in the mid-20th century. At the end of Tiberius' reign, Christianity had spread to the gentiles as a result of the preaching of St. Paul in Anatolia and in Greece. At the same time, Christianity continued to make progress among the Jews of Jerusalem, Alexandria, and Syria and quickly reached even Osroëne and the Parthian towns of the Euphrates, where Jewish colonies were numerous. The Roman authorities at first had difficulty in distinguishing the "Christos" believers from the orthodox Jews, but the religion of the former, on leaving its original milieu, quickly became differentiated.

A familiar charge against the Jews, however, continued to pursue the Christians: that they felt a hatred of mankind. Their expectation of the end of the world aroused a suspicion that that was what they indeed desired; moreover, they were also suspect for their aloofness—they cut themselves off from family and community—and for their meetings, whose purpose was obscure. Their 2nd-century spokesmen had to dispel the belief, often recorded, that they practiced magic involving cannibalism; further, that they indulged in sex orgies, incestuous to boot; and, the most common accusation of all, that they were atheists—people who denied the existence of the gods and rejected accepted cults. This last charge, which was, of course, exactly on the mark, must be set in the context of occasional episodes of mob violence against (non-Christian) atheists or doubters. Here the association of Christians with Jews, equally monotheistic, might have provided some protection for the Christians, but the Jews were faithful to a cult of the greatest antiquity and, moreover, had long made their peace with Caesar, Augustus, and their successors. It was a peace that could not extend to people who had (it would be alleged) apostasized from their own Judaism. Christians did not participate in the Jewish revolt of 66–73, and, under the Flavians, Christianity completely severed itself from its origins.

At this time the East was the centre of the new religion, whose followers grew in numbers from Egypt to the Black Sea and were beginning to be noticed in Bithynia and in Greece. Christians seemed fairly numerous in Rome as early as the end of the 1st century. When the age of the Apostles ended, the age of the church began, with its bishops, presbyters, and deacons, with its catechism, preaching, and celebration of the Eucharist. In the 2nd century, Christianity began to reach the intellectuals. Hellenistic culture offered educated Christians the resources of philosophical dialectic and of sophist rhetoric. The ex-

Beginnings of Christian church organization

ample of Philo of Alexandria had shown in the 1st century that it was possible to reconcile the Bible with the great Platonic ideas. By the 2nd century the Christian "apologists" tried to show that Christianity was in harmony with Greco-Roman humanism and that it was intellectually, and above all morally, superior to paganism.

But the Christians did not succeed in convincing the authorities. The first persecution, that of Nero, was related to a devastating fire in the capital in 64, for which the Christians were blamed or, perhaps, only made the scapegoats. In any case, their position as bad people (*mali homines* of the sort a governor should try to suppress) had been established, and later suppressions could be justified by reference to "the Neronian practice." So far as criminal law was concerned, such a precedent had considerable authority, of the sort that Pliny, as governor, was looking for in his handling of the Christians of Bithynia-Pontus in 111. His master, the emperor Trajan, told him not to seek them out but to execute those who, being informed against, refused to abjure their religion. Hadrian and other successors hewed to the same line thereafter. Thus, the persecutions remained localized and sporadic and were the result of private denunciations or of spontaneous popular protests. Under Marcus Aurelius, the difficulties of the times often caused the Christians, who refused to sacrifice to the state gods and to participate in the imperial cult, to be accused of provoking the wrath of the gods: martyrs appeared in the East, in Rome, in Gaul, and in Africa. Commodus' reign was more favourable to them, perhaps because certain members of his circle, not a very edifying one in other respects, were Christians or Christian sympathizers. This reprieve, however, was short-lived: Septimius Severus inaugurated the first systematic persecution. In 202 an edict forbade Christian (and Jewish) proselytism. Members of extremist sects were persecuted for preaching continence (which violated Augustus' laws against celibacy), for holding the state in contempt, and especially for refusing military service. Under Caracalla, the situation quieted, and the church continued to progress, favoured perhaps by the relative freedom that the law granted to funerary *collegia* (whence the first catacombs).

Cultural life from the Antonines to Constantine. Latin literature enjoyed its "Silver Age" under the Antonines, with the majority of great authors, such as Tacitus, Juvenal, and Pliny the Younger, having begun their careers under Domitian. They had no heirs: after Tacitus, Roman history was reduced to biography. It was only in the 4th century that history began to flourish again, with Ammianus Marcellinus, a Greek writing in Latin. Satire, the Roman genre par excellence, came to an end with Juvenal; and Pliny the Younger, a diligent rhetorician but with a lesser degree of talent, had only the mediocre Fronto as a successor. More original was the aforementioned rhetorician, scholar, and picaresque novelist Apuleius of Madauros.

Decline of Latin literature

A Greek renaissance, however, took place during the 2nd century. The Second Sophistic school reigned in every area: in rhetoric, history, philosophy, and even in the sciences. Schools of rhetoric and philosophy prospered in the East—in Smyrna, Ephesus, Pergamum, Rhodes, Alexandria, and even in Athens—protected and subsidized by the emperors, from Vespasian to Marcus Aurelius. The great sophists were Herodes Atticus, a multimillionaire from Athens; Polemon; and Aelius Aristides, a valetudinarian devotee of Asclepius. Dio Cassius and Herodian were conscientious and useful historians (first half of the 3rd century), as was later Dexippus the Athenian, whose work survives only in fragments. Science was represented by the mathematician Nicomachus of Gerasa; medicine, by Galen of Pergamum; astronomy, by the Alexandrian Ptolemy. Law remained the only Roman science, exemplified under the Antonines by Salvius Julianus and Gaius (the *Institutiones*) and rising to its zenith in the 3rd century as a result of the works of three jurists: Papinian, Ulpian, and Modestinus. Philosophy, heavily influenced by rhetoric and ethics, was represented under Domitian and Trajan by Dio (or Chrysostom) of Prusa, who outlined the stoical doctrine of the ideal sovereign. The biographer Plutarch and Lucian of Samosata were more eclectic, especially Lucian, who resembled Voltaire in his caustic skepticism. Under Marcus Aurelius, one of Lucian's friends, Celsus, wrote the first serious criticism of Christianity, "The True Word," known through Origen's refutation of it in the 3rd century. At this time philosophy leaned toward religious mysticism: under the Severans, Ammonius Saccas created the school of Alexandria, and his disciple Plotinus founded the Neoplatonist school, which was to fight bitterly against Christianity. After the apologists and, above all, Tertullian (*c.* 160–after 222), Christian thought deepened, and theology made its appearance. Clement and Origen (*c.* 185–*c.* 254), the greatest theologian of the time, were the luminaries of the church of Alexandria; the Roman church still wrote in Greek and was represented by the slightly old-fashioned Hippolytus; and the church of Africa had a powerful personality, St. Cyprian, bishop of Carthage.

The Christian theologians

The disappearance of the great lyric and poetic styles, the fossilizing of education as it came to be completely based on rhetoric (*paideia*), and the growing importance of philosophical and religious polemical literature among both pagans and Christians were the basic traits that, as early as the 3rd century, foreshadowed the intellectual life of the late empire.

Military anarchy and the disintegration of the empire (235–270). *Succession of emperors and usurpers.* The period from the death of Severus Alexander to the time of Claudius II Gothicus was marked by usurpations and barbarian invasions. After Maximinus the Thracian, who bravely fought the Alemanni but showed great hostility toward the Senate and the educated elite, the Gordians rose to power as a result of a revolt by wealthy African landowners. A senatorial reaction first imposed civilian emperors, Pupienus and Balbinus together, and then named Gordian III, a youth backed by his father-in-law, the praetorian prefect Timesitheus. Gordian III was murdered by the soldiers during a campaign against the Persians and was replaced, first by Philip the Arabian and then by Decius, both soldiers. Decius tried to restore Roman traditions and also persecuted the Christians, but he was killed by the Goths in 251 in a battle near the Black Sea. From 253 to 268 two Roman senators, Valerian and his son Gallienus, reigned. Valerian revived the persecution of the Christians, but he was captured by the Persians during a disastrous campaign and died in captivity (260). His son then reigned alone, facing multiple invasions and several usurpations. He moved constantly between the Rhine and the Danube, achieving brilliant victories (Milan in 262, the Nestus in 267), but the Pannonian army raised several competitors against him (Ingenuus, Regalianus, Aureolus). Too busy to protect the Gauls against the Franks and the Alemanni and the East against the Persians, he had to tolerate the formation of the Gallic empire under the praetorian prefect Marcus Cassianius Postumus (259–268) and the Palmyrene kingdom of Odenathus (260–267). Some of his reforms were a foreshadowing of the future: the senators were practically excluded from the army; the equites received the majority of commands and of provincial governorships; and the composition of the army was modified by the creation of new army corps and especially of a strong cavalry, which was placed under the command of a single leader and charged with closing the breaches that the barbarians were opening along the frontiers. Upon his father's death Gallienus had put an end to the persecution of the Christians, preferring to fight the new religion through intellectual means; to that end, he favoured the ancient Greek cults (Demeter of Eleusis) and protected the Neoplatonist philosopher Plotinus. These initiatives increased the number of his enemies, particularly among the patriotic senators and the Pannonian generals. While Gallienus was in Milan besieging the usurper Aureolus, he was killed by his chiefs of staff, who proclaimed Claudius II (268), the first of the Illyrian emperors. The new emperor won a great victory against the Alemanni on the Garda lake and overwhelmed the Goths in Naissus (269) but died of the plague in 270. This fatal period brought to light one of the major defects of the empire: the lack of a legitimate principle of succession and the preponderant role of the army in politics. The structures that had

Victories over the Germans

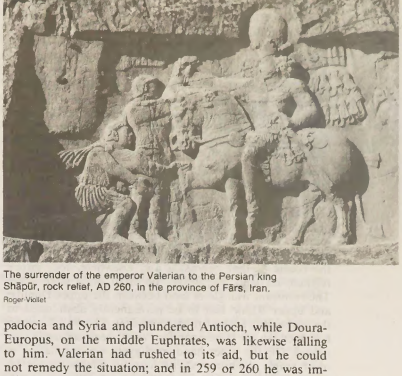
created the strength of the principate were weakened, and the empire required deep reforms. Gallienus had felt their necessity but had been too weak to impose them.

The barbarian invasions. The Goths were Germans coming from what is now Sweden and were followed by the Vandals, the Burgundians, and the Gepidae. The aftereffect of their march to the southeast, toward the Black Sea, was to push the Marcomanni, the Quadi, and the Sarmatians onto the Roman *limes* in Marcus Aurelius' time. Their presence was brusquely revealed when they attacked the Greek towns on the Black Sea about 238. Timesitheus fought against them under Gordian III, and under Philip and Decius they besieged the towns of Moesia and Thrace, led by their kings, Ostrogotha and Kniva. Beginning in 253, the Crimean Goths and the Heruli appeared and dared to venture on the seas, ravaging the shores of the Black Sea and the Aegean as well as several Greek towns. In 267 Athens was taken and plundered despite a strong defense by the historian Dexippus. After the victories of Gallienus on the Nestus and Claudius at Naissus (Nish), there was for a time less danger. But the countries of the middle Danube were still under pressure by the Marcomanni, Quadi, Iazyges, Sarmatians, and the Carpi of free Dacia, who were later joined by the Roxolani and the Vandals. In spite of stubborn resistance, Dacia was gradually overwhelmed, and it was abandoned by the Roman troops, though not evacuated officially. When Valerian was captured in AD 259/260, the Pannonians were gravely threatened, and Regalianus, one of the usurpers proclaimed by the Pannonian legions, died fighting the invaders. The defense was concentrated around Sirmium and Siscia-Poetovio, the ancient fortresses that had been restored by Gallienus, and many cities were burned.

In the West the invasions were particularly violent. The Germans and the Gauls were driven back several times by the confederated Frankish tribes of the North Sea coast and by the Alemanni from the middle and upper Rhine. Gallienus fought bitterly, concentrating his defense around Mainz and Cologne, but the usurpations in Pannonia prevented him from obtaining any lasting results. In 259–260 the Alemanni came through the Agri Decumates (the territory around the Black Forest), which was now lost to the Romans. Some of the Alemanni headed for Italy across the Alpine passes; others attacked Gaul, devastating the entire eastern part of the country. Passing through the Rhône Valley, they eventually reached the Mediterranean; and some bands even continued into Spain. There they joined the Franks, many of whom had come by ship from the North Sea, after having plundered the western part of Gaul. Sailing up the estuaries of the great rivers, they had reached Spain and then, crossing the Strait of Gibraltar, had proceeded to Mauretania Tingitana. Gallienus, outflanked, entrusted Gaul and his young son Saloninus to Postumus, who then killed Saloninus and proclaimed himself emperor. The several invasions had so frightened the people that the new emperor was readily accepted, even in Spain and Britain. He devoted himself first to the defense of the country and was finally considered a legitimate emperor, having established himself as a rival to Gallienus, who had tried in vain to eliminate him but finally had to tolerate him. Postumus governed with moderation, and, in good Roman fashion, minted excellent coins. He, too, was killed by his soldiers, but he had successors who lasted until 274.

Alemanni invasions

Difficulties in the East. In the East the frontiers had been fixed by Hadrian at the Euphrates. But under Nero, the Romans had claimed control over the kings of Armenia, and under Caracalla they had annexed Osroëne and Upper Mesopotamia. The Parthian empire had been weak and often troubled, but the Sāsānids were more dangerous. In 241, Shāpūr I (Sapor), an ambitious organizer and statesman, mounted the throne: he united his empire by bringing the Iranian lords into line and by protecting the Zoroastrian religion. He also tolerated the Manichaeans and put an end to the persecutions of the Christians and Jews, thereby gaining the sympathy of these communities. In 252, with a large army at his command, Shāpūr imposed Artavasdes on Armenia, attacked Mesopotamia, and took Nisibis. In 256 his advance troops entered Cappadocia and Syria and plundered Antioch, while Dura-Europus, on the middle Euphrates, was likewise falling to him. Valerian had rushed to its aid, but he could not remedy the situation; and in 259 or 260 he was imprisoned by Shāpūr during operations about which little is known. Mesopotamia was lost and Rome was pushed back to the Euphrates. Cappadocia, Cilicia, and Syria were again plundered, and a puppet emperor was appointed in Antioch. But these victories were transitory: in Osroëne, Edessa had shown resistance, a defense was organized in Cappadocia and Cilicia, and Odenathus, the prince of Palmyra, took Shāpūr by surprise and forced him back to Iran. Having thus aided the Roman cause, Odenathus then began to act in his own interest: he continued the fight against the Persians and took the title "King of Kings." The Romans officially entrusted him with the defense of the East and conferred on him the governorship of several provinces; the "kingdom" of Palmyra thus extended from Cilicia to Arabia. He was murdered in 267 without ever having severed his ties with Gallienus. His widow Zenobia had her husband's titles granted to their son Vaballathus. Then in 270, taking advantage of the deaths of Gallienus and Claudius II, she invaded Egypt and a part of Anatolia. This invasion was followed by a rupture with Rome, and in 271 Vaballathus was proclaimed Imperator Caesar Augustus. The latent separatism of the Eastern provinces and, undoubtedly, some commercial advantages caused them to accept Palmyrene domination without difficulty, as they had, in the past, supported Avidius Cassius and Pescennius Niger against the legitimate emperors. In 272 unity was restored by Aurelian, but Mesopotamia was lost, and the Euphrates became the new frontier of the empire.



The surrender of the emperor Valerian to the Persian king Shāpūr, rock relief, AD 260, in the province of Fārs, Iran.
Roger-Viollet

Economic and social crisis. The invasions and the civil wars worked in combination to disrupt and weaken the empire over a span of half a century. Things were at their worst in the 260s, but the entire period from 235 to 284 brought the empire close to collapse. Many regions were laid waste (northern Gaul, Dacia, Moesia, Thrace, and numerous towns on the Aegean); many important cities had been pillaged or destroyed (Byzantium, Antioch, Olbia, Lugdunum); and northern Italy (Cisalpine Gaul) had been overrun by the Alemanni. During the crisis, the emperor either focused his forces on the defense of one point, inviting attack at another, or he left some embattled frontier altogether to its own devices; any commander who proved successful had the emperorship thrust upon him, on the very heels of his victories over the invaders. Counting several sons and brothers, more than 40 emperors thus established themselves for a reign of some sort, long or (more often) short. The political destabilization fed on itself, but it also was responsible for heavy expenditure of life and treasure. To keep pace with the latter, successive emperors rapidly and radically reduced the percentage of precious metal in the standard silver coins to almost nothing so as to spread it over larger issues. What thus became

Political destabilization and economic damage

a fiduciary currency held up not too badly until the 260s, when confidence collapsed and people rushed to turn the money they had into goods of real value. An incredible inflation got under way, lasting for decades.

The severity of damage done to the empire by the political and economic destabilization is not easily estimated since for this period the sources of every sort are extremely poor. Common sense would suggest that commerce was disrupted, taxes collected more harshly and unevenly, homes and harvests destroyed, the value of savings lost to inflation, and the economy in general badly shaken. A severe plague is reported that lasted for years in mid-century, producing terrible casualties. In some western areas, archaeology provides illustration of what one might expect: cities in Gaul were walled, usually in much reduced circuits; villas here and there throughout the Rhine and Danube provinces also were walled; road systems were defended by lines of fortlets in northern Gaul and adjoining Germany; and a few areas, such as Brittany, were abandoned or relapsed into pre-Roman primitiveness. Off the coasts of that peninsula and elsewhere, too, piracy reigned; on land, brigandage occurred on a large scale. The reentrant triangle of land between the upper Danube and upper Rhine had to be permanently abandoned to the barbarians around it in about 260. The Pax Romana had then, in all these manifest ways, been seriously disrupted. On the other hand, in Egypt, where inflation is most amply documented, its harmful effects cannot be detected. The Egyptian economy showed no signs of collapse. Furthermore, some regions—most of Britain, for example—emerged from the half-century of crisis in a more prosperous condition than before. A summary of the effects of crisis can only underline one single fact that is almost self-evident: the wonders of civilization attained under the Antonines required an essentially political base. They required a strong, stable monarchy in command of a strong army. If either or both were seriously disturbed, the economy would suffer, along with the civilization's ease and brilliance. If, on the other hand, the political base could be restored, the health of the empire as a whole was not beyond recovery.

In the meantime, certain broad changes unconnected with the political and economic crisis were going forward in the 3rd century. Civilians increasingly complained of harassment and extortion by troops stationed among them; exaction of taxes intended for the army also became the target of more frequent complaint; and demands by soldiers to interfere in civilian government, foremost by those stationed in the capital, grew more insolent. The choice of emperor became more and more openly the prerogative of the military, not the Senate; and, as mentioned, in the 260s senators were being largely displaced from high military commands. The equestrian rank, in which persons risen from military careers were often to be found, was the beneficiary of the new policy. In sum, the power of the military, high and low, was asserting itself against that of the civilians. From this change, further, there flowed certain cultural consequences; for, continuing the tendencies detectable even in the 1st century, the army was increasingly recruited from the most backward areas, above all, from the Danubian provinces. Here, too—indeed, throughout the whole northern glacis of the empire—it had been state policy to allow entire tribes of barbarians to immigrate and to settle on vacant lands, where they dwelled, farmed, paid taxes, and offered their sons to the army. Such immigrants, in increasingly large numbers from the reign of Marcus Aurelius on, produced, with the rural population, a very non-Romanized mix. From the midst of just such people, Maximinus mounted to the throne in 235, and later, likewise, Galerius (Caesar from 293). It is quite appropriate aesthetically, from Aurelian on, that these later 3rd-century rulers chose to present themselves to their subjects in their propaganda with stubbly chin, set jaw, and close-cropped hair on a bullet head.

The recovery of the empire and the establishment of the dominate (270–337). *The Illyrian emperors.* After Claudius II's unexpected death, the empire was ruled from 270 to 284 by several "Illyrian" emperors, who were good generals and who tried in an energetic way to restore equilibrium. The most remarkable was Aurelian. He first gained hard-won victories over the Alemanni and the Juthungi, who had invaded the Alpine provinces and northern Italy. To cheer the inhabitants of Rome, who had succumbed to panic, he began construction of the famous rampart known as Aurelian's Wall. And while crossing the Danubian provinces, before marching against Palmyra, he decided on an orderly evacuation of Dacia, an undefendable region that had been occupied by the barbarians since the time of Gallienus. In the East, he defeated Zenobia's troops easily and occupied Palmyra in 272. Shortly afterward, an uprising broke out in Egypt under the instigation of a rich merchant, who, like a great part of the population, was a partisan of the Palmyrene queen. In response, Aurelian undertook a second campaign, plundering Palmyra and subjugating Alexandria. These troubles, however, along with the devastation of the great caravan city, were to set back Roman trade seriously in the East. Later, rounding back on the Gallic empire of Postumus' successors, he easily defeated Tetricus, a peaceful man not very willing to fight, near Cabillonum. The unity of the empire was restored, and Aurelian celebrated a splendid triumph in Rome. He also reestablished discipline in the state, sternly quelled a riot of artisans in the mints of Rome, organized the provisioning of the city by militarizing several corporations (the bakers, the pork merchants), and tried to stop the inflation by minting an antoninianus of sounder value. His religious policy was original: in order to strengthen the moral unity of the empire and his own power, he declared himself to be the protégé of the Sol Invictus (the Invincible Sun) and built a magnificent temple for this god with the Palmyrene spoils. Aurelian was also sometimes officially called *dominus et deus:* the principate had definitely been succeeded by the "dominate." In 275 Aurelian was murdered by certain officers who mistakenly believed that their lives were in danger. Aurelian's conquests

For once, his successor, the aged senator Tacitus, was chosen by the Senate—at the army's request and on short notice; he reigned only for a few months. After him, Probus, another Illyrian general, inherited a fortified empire but had to fight hard in Gaul, where serious invasions occurred in 275–277. Thereafter, Probus devoted himself to economic restoration; he attempted to return abandoned farmland to cultivation and, with the aid of military labour, undertook works of improvement. To remedy the depopulation, he admitted to the empire, as had Aurelian, a great number of defeated Goths, Alemanni, and Franks and permitted them to settle on plots of land in Gaul and in the Danubian provinces. After the assassination of Probus in 282 by soldiers, Carus became emperor and immediately associated with himself his two sons, Carinus and Numerian. Carus and Numerian fought a victorious campaign against the Persians but died under unknown circumstances. Carinus, left behind in the West, was later defeated and killed by Diocletian, who was proclaimed emperor in November 284 by the army of the East.

Diocletian. Diocletian may be considered the real founder of the late empire, though the form of government he established—the tetrarchy, or four persons sharing power simultaneously—was transitory. His reforms, however, lasted longer. Military exigencies, not the desire to apply a preconceived system, explain the successive nomination of Maximian as Caesar and later as Augustus in 286 and of Constantius and Galerius as Caesars in 293. The tetrarchy was a collegium of emperors comprising two groups: at its head, two Augusti, older men who made the decisions; and, in a secondary position, two Caesars, younger, with a more executive role. All four were related either by adoption or by marriage, and all were Illyrians who had attained high commands after a long military career. Of the four, only Diocletian was a statesman. The unity of the empire was safeguarded, despite appearances, for there was no territorial partitioning. Each emperor received troops and a sector of operation: Maximian, Italy and Africa; Constantius, Gaul and Britain; Galerius, the Danubian countries; and Diocletian, the East. Practically all governmental decisions were made by Diocletian, from The tetrarchy



Diocletian's tetrarchy; Diocletian and Maximian are at right, Constantius and Galerius in the front, in red porphyry, *c.* AD 300, brought to Venice in 1258.
Alinari/Art Resource, New York City

whom the others had received their power. He legislated, designated consuls, and retained precedence. After 287 he declared his kinship with the god Jupiter (Jove), who Diocletian claimed was his special protector. Diocletian, together with his Caesar Galerius, formed the "Jovii" dynasty, whereas Maximian and Constantius, claiming descent from the mythical hero Hercules, formed the "Herculii." This "Epiphany of the Tetrarchs" served as the divine foundation of the regime. The ideological recourse to two traditional Roman divinities represented a break with the Orientalizing attempts of Elagabalus and Aurelian. Even though he honoured Mithra equally, Diocletian wanted to be seen as continuing the work of Augustus. In dividing power, Diocletian's aim was to avoid usurpations, or at least to stifle them quickly—as in the attempt of Carausius, chief of the army of Britain, who was killed (293), as was his successor, Allectus (296), after a landing by Constantius.

The deification of the imperial function, marked by elaborate rituals, tended to set the emperors above the rest of mankind. But it was still necessary to avoid future rivalries and to assure the tetrarchy a legitimate and regular succession. Some time between 300 and 303 Diocletian found an original solution. After the anniversary of their 20-year reign the two Augusti abdicated (Maximian quite unwillingly), and on the same day (May 1, 305) the two Caesars became Augusti. Two new Caesars were chosen, Severus and Maximinus Daia, both friends of Galerius, whose strong personality dominated Constantius. In repudiating the principle of natural heredity (Maximian and Constantius each had an adult son), Diocletian took a great risk: absolute divine monarchy, which Diocletian largely established, implies the hereditary transmission of power, and the future was soon to demonstrate the attachment of the troops and even of the population to the hereditary principle.

Changes in provincial organization

In order to create a more efficient unity between subjects and administrators, Diocletian multiplied the number of provinces; even Italy was divided into a dozen small units of the provincial type. Rome, moreover, was no longer the effective capital of the empire, each emperor having his own residence in the part of the empire over which he ruled (Trier, Milan, Sirmium, Nicomedia). Although a few provinces were still governed by senators (proconsuls or consuls), the majority were given to equestrian *praesides*, usually without any military power but with responsibility for the entirety of civil administration (justice, police, finances, and taxes). The cities lost their autonomy, and the curiales administered and collected the taxes under the governor's direct control. The breaking up of the provinces was compensated for by their regrouping into a dozen dioceses, under equestrian vicars who were responsible to the emperor alone. The two praetorian prefects had less military power but played an important role in legislative, judicial, and above all, financial matters: the administration of the *annona*, which had become the basis of the fiscal system, in fact gave them management of the entire economy. Within the central administration the number of offices increased, their managers being civilians who carried out their functions as a regular career. All officials were enrolled in the *militia*, whose hierarchy was to be outlined during the 4th century.

Great efforts were devoted to strengthening the borders, and the *limes* were outfitted with fortresses (*castella*) and small forts (*burgi*), notably in Syria. The army's strength was increased to 60 legions (but with reduced personnel); and, in principle, each border province received a garrison of two legions, complemented by subsidiary troops. Adopting one of Gallienus' ideas, Diocletian created an embryonic tactical army under the direct orders of the emperor whose escort (*comitatus*) it formed. The troops were most often commanded by *duces* and *praepositi* rather than by provincial governors and were mainly recruited from among the sons of soldiers and from barbarians who enlisted individually or by whole tribes. In addition, the landowners had to provide either recruits or a corresponding sum of money. All of these reforms were instituted gradually, during defensive wars whose success demonstrated the regime's efficiency. Constantius put down Carausius' attempted usurpation and fought the Alemanni fiercely near Basel; Maximian first hunted down the Bagaudae (gangs of fugitive peasant brigands) in Gaul, then fought the Moorish tribes in Africa, in 296–298, triumphing at Carthage; and on the Danube, Diocletian, and later Galerius, conquered the Bastarnae, the Iazyges, and the Carpi, deporting them in large numbers to the provinces. In the East, however, the opposition of the Persians, led by the enterprising Narses, extended from Egypt to Armenia. The Persians incited uprisings by both the Blemmyes nomads in southern Egypt and the Saracens of the Syrian desert and made use of anti-Roman propaganda by the Manichaeans and Jews. Diocletian succeeded in putting down the revolt in Egypt and fortified the south against the Blemmyes. But in 297, Narses, the heir to Shāpūr's ambitions, precipitated a war by taking Armenia, Osroëne, and part of Syria. After an initial defeat, Galerius won a great victory over Narses, and in 298 the peace of Nisibis reinstated a Roman protégé in Armenia and gave the empire a part of Upper Mesopotamia that extended even beyond the Tigris. Peace was thus assured for some decades.

Persian opposition

The wars, the reforms, and the increase in the number of officials were costly, and inflation reduced the resources of the state. The *annona*, set up by Septimius Severus, had proved imperfect, and Diocletian now reformed it through the *jugatio-capitatio* system: henceforth, the land tax, paid in kind by all landowners, would be calculated by the assessment of fiscal units based on extent and quality of land, type of crops grown, number of settlers and cattle, and amount of equipment. The fiscal valuation of each piece of property, estimated in *juga* and *capita* (interchangeable terms whose use varied by region and period of time), required a number of declarations and censuses similar to those practiced long before in Egypt. Each year, the government established the rate of tax per fiscal unit; and every 15 years, beginning in 312, taxes were reassessed. This complicated system was not carried out uniformly in every region. Nevertheless, it resulted in an improved accounting of the empire's resources and a certain progress in fiscal equity, thus making the administration's heavy demands less unbearable. In addition, Diocletian wished to reorganize the coinage and stabilize

inflation. He thus minted improved sterling coins and fixed their value in relation to a gold standard. Nevertheless, inflation again became disturbing by the end of the century, and Diocletian proclaimed his well-known Edictum de Maximis Pretiis, fixing price ceilings for foodstuffs and for goods and services, which could not be exceeded under pain of death. The edict had indifferent results and was scarcely applied, but the inscriptions revealing it have great economic interest.

Late Roman society

Diocletian's reforms adumbrated the principal features of late Roman society: a society defined in all parts that could be useful to the state by laws fixing status and, through status, responsibility. The persons owning grain mills in Rome were (to anticipate developments that continued to unfold throughout the next two or three generations) responsible for the delivery of flour for the dole and could not bequeath or withdraw any part of their capital from their enterprise. Several other labour groups were similarly restricted, such as owners of seagoing vessels that served the supply of Rome, bargees in the Tiber, Ostian grain handlers, distributors of olive oil and pork for the dole, bath managers, and limeburners. A ban on moving to some other home or job along with production quotas were placed on people in trades serving state factories that made imperial court and army garments, cavalry equipment, and arms. Diocletian built a number of such factories, some in his capital Nicomedia, others in cities close to the groups whose needs they served. The laws imposing these obligations affected only labour groups serving the army and the capital (or capitals, plural, after the promotion of Constantinople); and, to identify them, induce them to serve, and hold them in their useful work, emperors as early as Claudius had offered privileges and imposed controls. Diocletian, however, greatly increased the weight and complexity of all these obligations.

Diocletian also changed the administrative districts in Egypt, in keeping with the model found elsewhere, by designating in each a central city to take responsibility for the whole. The last anomalous province was thus brought into line with the others. Everywhere, the imperial government continued to count on the members of the municipal senate to serve it, above all in tax collection but also in the supply of recruits, in rural police work, billeting for troops, or road building. As had been the case for centuries, they had to have a minimum of landed property to serve as surety for the performance of their administrative duties as well as to submit to nomination as senator, if it was so determined by the Senate. There had never been any one law to that effect, but by Diocletian's time the emperor had at his command a body of long-established custom and numerous imperial decisions that served just as well. Local elites were thus hereditary, compulsory agents of his purpose, exactly like the Tiber bargees.

Two other groups were frozen into their roles in the same fashion: soldiers and farmers. The sons of soldiers were required to take up their fathers' occupation (a law to that effect was in operation at least by 313); and the natural tendency of tenant farmers (*coloni*) to renew their lease on land that they, and perhaps their fathers and grandfathers, had worked was confirmed by imperial decisions—to such effect that, in 332, Constantine could speak of tenants on his Sardinian estates as bound to the acres they cultivated. This is the earliest explicit pronouncement on what is called the "colonate." Soon the institution was extended beyond imperial estates to tie certain categories of tenants to private estates as well. The emperors wanted to ensure tax revenue and, for that, a stable rural labour supply.

The colonate

The empire, as it is seen in abundant legislation for the period of Diocletian and beyond into the 5th century, has been called a "military dictatorship" or even a sort of totalitarian prison, in which every inhabitant had his own cell and his own shackles. This may well have been the rulers' intent. By their lights, such a system was needed to repair the weaknesses revealed in the 3rd-century crisis. The principle of hereditary obligations was not, after all, so very strange, set against the natural tendencies of the economy and the practices that had developed in earlier, easier times. Yet Diocletian's intentions could not be fully realized, given the limits on governmental effectiveness.

Persecutions of Christians

After a period of initial indifference toward the Christians, Diocletian ended his reign by unleashing against them, in 303, the last and most violent of their persecutions. It was urged on him by his Caesar Galerius and prolonged in the East for a decade (until 311) by Galerius as Augustus and by other emperors. As in earlier persecutions, the initiative arose at the heart of government; some emperors, as outraged by the Christians as many private citizens, considered it their duty to maintain harmony with the gods, the *pax deorum*, by which alone the empire flourished. Accordingly, Decius and Valerian in the 250s had dealt severely with the Christians, requiring them to demonstrate their apostasy by offering sacrifice at the local temples, and for the first time had directly struck the church's clergy and property. There were scores of Christians who preferred death, though the great majority complied or hid themselves. Within a matter of months after he had begun his attacks, however, Decius had died (251), and the bloody phase of Valerian's attacks also lasted only months (259/260). His son Gallienus had issued an edict of tolerance, and Aurelian was even appealed to by the church of Antioch to settle an internal dispute. Christianity had now become open and established, thanks to the power of its God so often, it seemed, manifested in miraculous acts and to the firmness with which converts were secured in a new life and community. The older slanders—cannibalism and incest—that had troubled the Apologists in the 2nd century no longer commanded credence. A measure of respectability had been won, along with recruits from the upper classes and gifts of land and money. By the end of the 3rd century Christians actually predominated in some of the smaller Eastern towns or districts, and they were well represented in Italy, Gaul, and Africa around Carthage; all told, they numbered perhaps as many as 5 million out of the empire's total population of 60 million. Occasional meetings on disputed matters might bring together dozens of bishops, and it was this institution or phenomenon that the Great Persecutions sought to defeat. The progress of a religion that could not accept the religious basis of the tetrarchy and certain of whose members were imprudent and provocative, as in the incidents at Nicomedia (where a church was built across from Diocletian's palace), finally aroused Galerius' fanaticism. In 303–304 several edicts, each increasingly stringent, ordered the destruction of the churches, the seizure of sacred books, the imprisonment of the clergy, and a sentence of death for all those who refused to sacrifice to the Roman gods. In the East, where Galerius was imposing his ideas more and more on the aging Diocletian, the persecution was extremely violent, especially in Egypt, Palestine, and the Danubian regions. In Italy, Maximian, zealous at the beginning, quickly tired; and in Gaul, Constantius merely destroyed a few churches without carrying reprisals any further. Nevertheless, Christianity could no longer be eradicated, for the people of the empire and even some officials no longer felt the blind hatred for Christians that had typified previous centuries.

The seven competing emperors

Struggle for power. The first tetrarchy had ended on May 1, 305; the second did not last long. After Constantius died at Eboracum in 306, the armies of Britain and Gaul, without observing the rules of the tetrarchic system, had hastened to proclaim Constantine, the young son of Constantius, as Augustus. Young Maxentius, the son of Maximian (who had never wanted to retire), thereupon had himself proclaimed in Rome, recalled his father into service, and got rid of Severus. Thus, in 307–308 there was great confusion. Seven emperors had, or pretended to have, the title of Augustus: Maximian, Galerius, Constantine, Maxentius, Maximinus Daia, Licinius (who had been promoted Augustus in 308 by Galerius against Constantine), and, in Africa, the usurper Domitius Alexander.

This situation was clarified by successive eliminations. In 310, after numerous intrigues, old Maximian was killed by his son-in-law Constantine, and in the following year Alexander was slain by one of Maxentius' praetorian prefects. In 311 Galerius died of illness a few days after having admitted the failure of his persecutions by proclaiming an edict of tolerance. There remained, in the West, Constantine and Maxentius and in the East, Licinius and Maximinus Daia. Constantine, the best general, invaded Italy with

a strong army of faithful Gauls and defeated Maxentius near the Milvian Bridge, not far from Rome. While attempting to escape, Maxentius drowned. Constantine then made an agreement with Licinius, and the two rallied the Eastern Christians to their side by guaranteeing them religious tolerance in the Edict of Milan (313). This left Maximinus Daia, now isolated and regarded as a persecutor, in a weak position; attacked by Licinius near Adrianople, he fell ill and died soon afterward, in 313. This left the empire with two leaders, Constantine and Licinius, allied in outward appearances and now brothers-in-law as a result of Licinius' marriage to Constantine's sister.

The reign of Constantine. Constantine and Licinius soon disputed among themselves for the empire. Constantine attacked his adversary for the first time in 316, taking the dioceses of Pannonia and Moesia from him. A truce between them lasted 10 years. In 316 Diocletian died in Salona, which he had never felt a desire to leave despite the collapse of his political creation. Constantine and Licinius then reverted to the principles of heredity, designating three potential Caesars from among their respective sons, all still infants, with the intention of securing their dynasties (two sons of Constantine and one of Licinius). The dynastic concept, however, required the existence of only a single emperor, who imposed his own descendance. Although Constantine favoured the Christians, Licinius resumed the persecutions, and in 324 war erupted once again. Licinius, defeated first at Adrianople and then in Anatolia, was obliged to surrender and, together with his son, was executed. Next, Constantine's third son, Constantius, was in turn named Caesar, as his two elder brothers, Crispus and Constantine the Younger, had been some time before. The second Flavian dynasty was thus founded, and Constantine let it be believed that his father, Flavius Constantius (Chlorus), was descended from Claudius Gothicus.

Constantine's conversion to Christianity had a far-reaching effect. Like his father, he had originally been a votary of the Sun; worshiping at the Grand Temple of the Sun in the Vosges Mountains of Gaul, he had had his first vision—albeit a pagan one. During his campaign against Maxentius, he had had a second vision—a lighted cross in the sky—after which he had painted on his men's shields a figure that was perhaps Christ's monogram (although he probably had Christ confused with the Sun in his manifestation as *summa divinitas* ["the highest divinity"]). After his victory he declared himself Christian. His conversion remains somewhat mysterious and his contemporaries—Lactantius and Eusebius of Caesarea—are scarcely enlightening and even rather contradictory on the subject. But it was doubtless a sincere conversion, for Constantine had a religious turn of mind. He was also progressive and greatly influenced by the capable bishops who surrounded him from the very beginning.

Constantine's conversion to Christianity

Until 320–322 solar symbols appeared on Constantine's monuments and coins, and he was never a great theologian. Yet his favourable policy toward the Christians never faltered. Christianity was still a minority religion in the empire, especially in the West and in the countryside (and consequently within his own army), thus excluding the possibility of any political calculation on his part. But it was enthusiastically welcomed in the East, and thanks to Constantine the new religion triumphed more rapidly; his official support led to the conversion of numerous pagans, although with doubtful sincerity because they were indifferent in their moral conviction.

The church, so recently persecuted, was now suddenly showered with favours: the construction of magnificent churches (Rome, Constantinople), donations and grants, exemptions from decurial duties for the clergy, juridical competences for the bishops, and exceptional promotions for Christian officials. Pagans were not persecuted, however, and Constantine retained the title of *pontifex maximus*. But he spoke of the pagan gods with contempt and forbade certain types of worship, principally nocturnal sacrifices. In 331 he ordered an inventory of pagan property, despoiled the temples of their treasure, and finally destroyed a few Eastern sanctuaries on the pretext of immorality.

The churches were soon to feel the burden of imperial solicitude: the "secular arm" (*i.e.*, the government) was placed at the service of a fluctuating orthodoxy, for the emperor was impressionable to arguments of various coteries and became quite lost in theological subtleties. In 314 the Council of Arles had tried in vain to stop the Donatist schism (a nationalistic heretical movement questioning the worthiness of certain church officials) that arose in Africa after Diocletian's persecutions. The Arian heresy raised even more difficulties: Arius, an Alexandrian priest and disciple of Lucian of Antioch, questioned the dogma of the Trinity and of the Godhead of Christ, and his asceticism, as well as the sharpness of his dialectics, brought him many followers; he was convicted several times, but the disorders continued. Constantine, solicited by both sides and untroubled by doctrinal nuances that were, moreover, foreign to most believers in the West, wished to institute a universal creed; with this in mind he convened the general Council of Nicaea, or Nicene Council, in 325. He condemned Arius and declared, in spite of the Easterners, that Jesus was "of one substance" with God the Father. Nevertheless, the heresy continued to exist, for Constantine changed his mind several times; he was influenced by Arian or semi-Arian bishops and was even baptized on his deathbed, in 337, by one of them, Eusebius of Nicomedia.

The Arian heresy

Between 325 and 337 Constantine effected important reforms, continuing Diocletian's work. The division between the *limitanei* border troops and the tactical troops (*comitatenses* and imperial guard) led by *magistri militum* was clarified, and military careers became independent of civil careers. At the same time, however, he lodged an increasing number of troops in or next to cities, a process whose objective was ease and economy of supply; however, training and discipline were harder to enforce because of it, and the men hung about in idleness. It was also under Constantine that a barbarian commander in the Roman army attained a historical significance. He was Crocus the Alaman, who led the movement among the troops that resulted in Constantine's seizure of the rank of Augustus in 306 immediately after his father Constantius' death. A similar figure was the great commander Bonitus, a Frank, in the years 316–324; and Constantine credited his victories against Maxentius in 311–312 principally to his barbarian troops, who were honoured on the triumphal Arch of Constantine in Rome. In opposition to him, Licinius mustered drafts of Goths to strengthen his army. Goths were also brought in by Constantine, to the number of 40,000, it is said, to help defend Constantinople in the latter part of his reign, and the palace guard was thenceforward composed mostly of Germans, from among whom a great many high army commands were filled. Dependence on immigrants or first-generation barbarians in war was to increase steadily, at a time when conventional Roman troops were losing military value.

Constantine raised many equestrians to senatorial rank, having in his earlier reign the still rapidly increasing ranks of the civil service to fill—it was at least 50 times the size of the civil service under Caracalla—and having in his later reign a second senate to fill, in Constantinople (see below). A rapid inflation in titles of honour also took place. As a result of these several changes, the equestrian order ceased to have meaning, and a new nobility of imperial service developed. Constantine gave first rank in the central administration to the palace quaestor, the *magister officiorum*, and the counts of finance (*comes sacrarum largitionum, comes rei privatae*). The diocesan vicars were made responsible to the praetorian prefects, whose number was increased and whose jurisdictions were now vast territories: the prefectures of Gaul, Italy, Illyricum, and the East. The unification of political power brought with it a corresponding decentralization of administration.

In order to reorganize finances and currency, Constantine minted two new coins: the silver miliarensis and, most importantly, the gold solidus, whose stability was to make it the Byzantine Empire's basic currency. And by plundering Licinius' treasury and despoiling the pagan temples, he was able to restore the finances of the state. Even so, he still had to create class taxes: the *gleba* for

senators, and the *chrysargyre,* which was levied in gold and silver on merchants and craftsmen in the towns.

Founding of Constantinople

Constantine's immortality, however, rests on his founding of Constantinople. This "New Rome," established in 324 on the site of Byzantium and dedicated in 330, rapidly increased in population as a result of favours granted to immigrants. A large number of churches were also built there, even though former temples were not destroyed; and the city became the administrative capital of the empire, receiving a senate and proconsul. This choice of site was due not to religious considerations, as has been suggested, but rather to reasons that were both strategic (its proximity to the Danube and Euphrates frontiers) and economic (the importance of the straits and of the junction between the great continental road, which went from Boulogne to the Black Sea, and the eastern commercial routes, passing through Anatolia to Antioch and Alexandria). Constantine died on May 22, 337.

The Roman Empire under the 4th-century successors of Constantine. *The rule of Constantine's sons.* After some months of confusion, Constantine's three surviving sons (Crispus, the eldest son, had been executed in mysterious circumstances in 326), supported by the armies faithful to their father's memory, divided the empire among themselves and had all the other members of their family killed. Constantine II kept the West, Constantius the East, and Constans, the youngest brother, received the central prefecture (Italy, Africa, and Illyricum). In 340 Constantine II tried to take this away from Constans but was killed. For the next 10 years there was peace between the two remaining brothers, and Constans won acceptance for a religious policy favourable to the Niceans, whose leader, Athanasius, had received a triumph in Alexandria. In 350 a mutiny broke out in Autun; Constans fled but was killed in Lugdunum by Magnentius, a usurper who was recognized in Gaul, Africa, and Italy. Constantius went out to engage Magnentius, and the Battle of Mursa (351) left the two strongest armies of the empire—those of Gaul and of the Danube—massacred, thus compromising the empire's defense. Magnentius retreated after his defeat and finally committed suicide in 353.

The reign of Constantius

Thenceforth, Constantius reigned alone as Augustus, aided by a meddlesome bureaucracy in which mission deputies (*agentes in rebus*), informers, and spies played an important role. He named two Caesars in succession, his two young surviving cousins, Gallus in the East and Julian in Gaul. Constantius eventually had to get rid of Gallus, who proved incompetent and cruel and soon terrorized Antioch. Julian, however, was a magnificient success, a fact that aroused Constantius' jealousy and led to Julian's usurpation; for the latter was proclaimed Augustus, in spite of Constantius' opposition, at Lutetia in 361. Civil war was averted when Constantius died in November 361, leaving the empire to Julian, the last ruler of the Constantinian family (see below).

At the time of his death in 337 Constantine had been preparing to go to war against the Persians. This legacy weighed heavily on the shoulders of Constantius, a military incompetent when compared to the energetic Sāsānian king Shāpūr II. Nearly every year the Persians attacked and pillaged Roman territory; the Mesopotamian towns were besieged, and Nisibis alone resisted. There was a lull between 350 and 357, while Shāpūr was detained by troubles in the eastern regions of his own kingdom. The war resumed, however, and Mesopotamia was partly lost when the emperor had to leave in order to fight Julian. Constantius had fought Shāpūr conscientiously, but his generals were mediocre, except for Urisicinus, and he himself was clumsy. In the meantime, the Rhine and Danube were threatened frequently, because the troops had been withdrawn from there and sent to the East. Constantius, moreover, had made a mistake in sending Chnodomar, the Alemannic king, against Magnentius in 351, for his tribes had gone on to ravage Gaul. Julian, however, soon revealed himself to be a great military leader by winning several well-fought campaigns between 356 and 361, most notably at Strasbourg in 357, and by restoring approximately 70 plundered villages. His abandonment, in AD 358, of the district of Toxandria, roughly equivalent to modern Belgium, to its barbarian squatters, on condition of their defending it against other invaders, was no doubt a realistic decision. Constantius defeated the Quadi and the Goths on the Danube in 359, but court intrigues, Magnentius' usurpation, and the interminable war against the Persians allowed the barbarians to wreak great havoc.

Constantius was primarily interested in religious affairs. His interventions created a "caesaro-papism" that was unfavourable to the church, for after the Battle of Mursa the emperor had become violently Arian. The Christological problem had moved to the forefront. In 360 Constantius obtained a new creed by force from the Council of Constantinople, which, rejecting the notion of "substance" as too risky, declared only that the Son was like the Father and thus left the problem unresolved. Pagans as well as orthodox Nicaeans (Homoousians) and extremist Arians (Anomoeans) were persecuted, for in 356–357 several edicts proscribed magic, divination, and sacrifices and ordered that the temples be closed. But when Constantius visited Rome in 357, he was so struck by its pagan grandeur that he apparently suspended the application of these measures.

Constantius' Arianism

The reign of Julian. Julian, who had been spared because of his tender age from the family butchering in 337, had been brought up far from the court and was undoubtedly intended for the priesthood. Nevertheless, he had been allowed to take courses in rhetoric and philosophy at Ephesus and later at Athens; he developed a fondness for Hellenic literature, and he secretly apostatized around 351. When he became sole emperor at the end of 361, he proclaimed his pagan faith, ordered the restitution of the temples seized under Constantius, and freed all the bishops who had been banished by the Arians, so as to weaken Christianity through the resumption of doctrinal disputes. The religion he himself espoused was compounded of traditional non-Christian elements of piety and theology, such as might have been found in any fairly intellectual person in the preceding centuries, along with elements of Neoplatonism developed by Porphyry and Iamblichus of two or three generations earlier, and, finally, much of the organization and social ethic of the church. From Neoplatonism he learned the techniques of direct communication with the gods (theurgy) through prayer and invocation; from the church he adopted, as the church itself had adopted from the empire's civil organization, a hierarchy of powers: provincial, metropolitan, urban, with himself as supreme pontiff. His deep love of traditional higher culture, moreover, provoked his war on Christian intellectuals and teachers who, he protested, had no right to Homer or Plato. Many Christians both before and later concurred with him, being themselves troubled by the relation between Christianity and inherited literature and thought, steeped as both were in pagan beliefs.

In the latter part of his 18-month reign, Julian forbade Christians from teaching, began the rebuilding of the Temple at Jerusalem, restored many pagan shrines, and displayed an exaggerated piety. Whereas Constantine (and his sons to a lesser degree) had introduced a huge number of coreligionists into the upper ranks of the army and government, achieving a rough parity between the members of the two religions, Julian began to reverse the process. Within a short while Julian was successful enough in his undertaking to have aroused the fear and hatred of the Christians, who for a long time thought of him as the Antichrist.

In the political realm, Julian wished to return to the liberal principate of the Antonines—to a time before the reforms of Diocletian and Constantine, whom he detested. He put an end to the terrorism of Constantius' eunuchs and *agentes in rebus* and reduced the personnel and expenditures of the court, while he himself lived like an ascetic. In the provinces he lightened the financial burden on individuals by reducing the *capitatio,* and on cities, by reducing the *aurum coronarium* and restoring the municipal properties confiscated by Constantius. On the other hand, he increased the number of curiales by reinstating numerous clerks in an attempt to return the ancient lustre to municipal life. Thus, he earned the gratitude of pagan intellectuals, who were enamoured of the past of

free Greece; and Ammianus made him the central hero of his history.

Julian's attempt to defeat Persia

Taking up Trajan's dream, Julian wished to defeat Persia definitively by engaging the empire's forces in an offensive war that would facilitate a national reconciliation around the gods of paganism. But his army was weak—corrupted perhaps by large numbers of hostile Christians. After a brilliant beginning, he was defeated near Ctesiphon and had to retrace his steps painfully; he was killed in an obscure encounter on June 26, 363.

Julian's successor, Jovian, chosen by the army's general staff, was a Christian, but not a fanatic. He negotiated a peace with Shāpūr, by which Rome lost a good part of Galerian's conquests of 298 (including Nisibis, which had not surrendered) and abandoned Armenia. He also restored tolerance in religious affairs, for he neither espoused any of the heresies nor persecuted pagans. In February 364 he died accidentally.

The reign of Valentinian and Valens. Once again the general staff unanimously chose a Pannonian officer—Valentinian, an energetic patriot and, like Jovian, a moderate Christian—but he had to yield to the rivalry of the armies by dividing authority. Taking the West for himself, Valentinian entrusted the East to his brother Valens, an inexperienced man whom he raised to the rank of Augustus. For the first time the two parts of the empire were truly separate, except for the selection of consuls, in which Valentinian had precedence.

Although he served the state with dedication, Valentinian could be brutal, choleric, and authoritarian. His foreign policy was excellent: all the while he was fighting barbarians (the Alemanni in Gaul, the Sarmatians and Quadi in Pannonia) and putting down revolts in Britain and Africa (notably that of the Berber Firmus) with the aid of his top general, Theodosius the Elder, he was taking care to improve the army's equipment and to protect Gaul by creating a brilliant fortification. His domestic measures favoured the curiales and the lower classes: from then on, taxes would be collected exclusively by officials; the protection of the poor was entrusted to "defenders of the plebs," chosen from among retired high officials (*honorati*). Nevertheless, needs of state obliged him to accentuate social immobility, to reinforce corporation discipline and official hierarchization, and to demand taxes ruthlessly. At first he was benevolent to the Senate of Rome, supervised the provisioning of the city, and legislated in favour of its university, the nursery of officials (law of 370). But beginning in 369, under the influence of Maximin, the prefect of Gaul, he initiated a period of terror, which struck the great senatorial families. Meanwhile, religious peace reigned in the West, tolerance was proclaimed, and after some difficulty, Rome found a great pope in Damasus, who, beginning in 373, actively supported the new bishop of Milan, St. Ambrose, an ardent defender of orthodoxy.

Religious peace in the West, persecutions in the East

In the East, Valens, who was incapable and suspicious, had fallen under the influence of legists, such as the praetorian prefect Modestus. The beginning of Valens' reign was shadowed by the attempted usurpation of Procopius (365–366), a pagan relative of Julian's who failed and was killed by the army, which remained faithful to Valens. Modestus instituted harsh persecutions in Antioch of the educated pagan elite. Valens was a fanatic Arian, who exiled even moderate Nicaean bishops and granted to Arians favours that aroused violent reactions from the orthodox, whose power had increased in the East. Valens' policies made the East prey to violent religious passions.

On the Danube, Valens fought the Visigoths and made a treaty with their king, Athanaric, in 369; but in 375 the Ostrogoths and the Greutingi appeared on the frontiers, pushed from their home in southern Russia by the powerful Huns. In 376 Valens authorized the starving masses to enter Thrace; but, being exploited and mistreated by the officials, they soon turned to uncontrollable pillaging. Their numbers continually increased by the addition of new bands, until finally they threatened Constantinople itself. Valens sent for aid from the West, but without waiting for it to arrive he joined battle and was killed in the Adrianople disaster of 378, which to some critics foreshadowed the approaching fall of the Roman Empire.

The Goths, who were also stirring up Thrace and Macedonia, could no longer be driven out. The provinces subject to their pillaging soon included Pannonia farther up the Danube, where Gratian agreed with a cluster of three tribal armies to settle them as a unit under their own chiefs on vacant lands (380). By a far more significant arrangement of the same sort two years later, Theodosius assigned to the Goths a large area of Thrace along the Danube as, in effect, their own kingdom; there they enjoyed autonomy as well as a handsome subsidy from the emperor, exactly as tribes beyond the empire had done in previous treaties. They were expected to respond to calls on their manpower if the Roman army needed supplementing, as it routinely did. Although the Goths considered this treaty ended with Theodosius' death and resumed their lawless wanderings for a while, it nevertheless represented the model for subsequent ones, again struck with the Goths under their king Alaric (from 395; see below) and with later barbarian tribes. The capture of the empire had begun.

The reign of Gratian and Theodosius I. Following Valentinian's sudden death in 375, the West was governed by his son Gratian, then 16 years old, who had been given the title of Augustus as early as 367. The Pannonian army, rife with intrigue, quickly proclaimed Gratian's half-brother, Valentinian II, only four years old. The latter received Illyricum under his older brother's guardianship, and this arrangement satisfied everybody. Valentinian's advisers were executed: Maximian was sacrificed to the spite of the Senate, and Theodosius the Elder became the victim of personal jealousies. Gratian announced a liberal principate, supported in Gaul by the wealthy family of the Bordeaux poet Ausonius and in Rome by the Symmachi and the Nicomachi Flaviani, representatives of the pagan aristocracy. His generals defeated the Alemanni and the Goths on the Danube but arrived too late to save Valens.

On Jan. 19, 379, before the army, Gratian proclaimed Theodosius, the son of the recently executed general, as Eastern emperor. Theodosius was chosen for his military ability and for his orthodoxy (Gratian, extremely pious, had come under the influence of Damasus and Ambrose). The East was enlarged by the dioceses of Dacia and Macedonia, taken from Valentinian II. Gratian and Theodosius agreed to admit the Goths into the empire, and Gratian applied the policy also to the Salian Franks in Germany. Theodosius soon dominated his weak colleague and entered the battle for the triumph of orthodoxy. In 380 the Arians were relieved of their churches in Constantinople, and in 381 the Nicaean faith was universally imposed by a council whose canons established the authority of the metroplitan bishops over their dioceses and gave the bishop of the capital a primacy similar to that of the bishop of Rome.

Influx of Goths and Franks

In ecclesiastical affairs, the separation between East and West was codified. The Westerners bowed to this policy, satisfied with the triumph of orthodoxy. Gratian then permitted Ambrose and Damasus to deal harshly with the Arians, with the support of the state. Paganism also was hounded: following Theodosius' lead, Gratian refused the chief priesthood, removed the altar of Victory from the hall of the Roman Senate, and deprived the pagan priests and the Vestal Virgins of their subsidies and privileges. The pagan senators were outraged, but their protests were futile because Gratian was watched over by Ambrose.

This militantly orthodox policy aroused the displeasure of the pagans and of the Western Arians: thus, when Gratian left Trier for Milan, the army of Gaul and Britain proclaimed its leader, Maximus, in 383. He conquered Gaul without difficulty, and Gratian was killed in Lyons. Maximus, who, like Theodosius, was Spanish and extremely orthodox, was recognized by the latter. In the meantime, the third Augustus, Valentinian II, had taken refuge in Milan after suffering defeat in Pannonia. He was effectively under the domination of his mother, Justina, an Arian who sought support for her son among the Arians and pagans of Rome and even among the African Donatists (a Christian heresy). In 388 Maximus, after arriving in Italy, first expelled Valentinian and then prepared to attack Theodosius. The latter, accepting the inevitability of

Sole rulership of Theodosius

war, strengthened his resolve and gained several victories. Maximus was killed at Aquileia in 388, and thenceforth Theodosius ruled both West and East; he was represented in the East by his son Arcadius, an Augustus since 383. Valentinian II was sent to Trier, accompanied by the Frankish general Arbogast to control him.

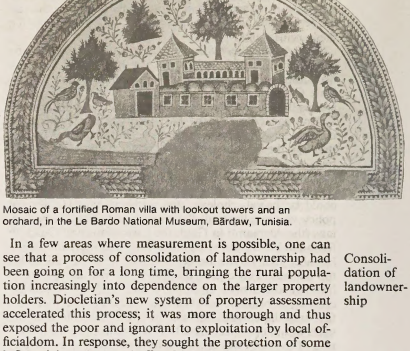
After a few years' respite, during the prefectureships of Nicomachus Flavianus in Rome and Tatian in the East, paganism waged its last fight: Theodosius, influenced by Ambrose, who had dared to inflict public penance on him in 390 after the massacre at Thessalonica, had determined to eliminate the pagans completely. After a few hostile clashes, the law of Nov. 8, 392, proscribed the pagan religion. Then Arbogast, after Valentinian II's death in 392 under shadowy circumstances, proclaimed as emperor the rhetorician Eugenius. When Theodosius refused to recognize him, Eugenius was thrown into the arms of the pagans of Rome. But this last "pagan reaction" was short-lived; in 394, with his victory at the Frigidus (modern Vipacco) River, between Aquileia and Emona, Theodosius put an end to the hopes of Eugenius and his followers. His intention was to place his son Honorius, proclaimed Augustus in 393, over the West, while returning his eldest son, Arcadius, to the East. But Theodosius' sudden death in January 395 precipitated the division of the empire.

Theodosius had successfully dealt with the danger of the Goths, although not without taking risks, and had both established a dynasty and imposed the strictest orthodoxy. A compromise peace with the Persians had given Rome, in 387, a small section of Armenia, where he had founded Theodosiopolis (Erzurum). He had survived two pretenders in the West. These military successes were, however, won with armies in which barbarians were in the majority, which was not a good sign. The barbarian presence is reflected in the names of his commanding officers, including such Franks as Richomer, Merovech, and Arbogast, and the half-Vandal Stilicho, who through his marriage to Serena, Theodosius' niece, had entered the imperial family.

Social and economic conditions. During the 4th century the emperor's power was theoretically absolute, the traditions of the principate having given way to the necessities of defense.

The emperor was both heir to the Hellenistic *basileus* (absolute king) and the anointed of the deity. Pagans and Christians alike considered him "emperor by the grace of God," which, strictly speaking, rendered the imperial cult unnecessary. Indeed, he hardly needed the ceremonies and parade of god-awfulness with which Diocletian and his successors were surrounded. Yet imperial authority had actually lost much of its effectiveness due to the growth and nature of late Roman government. Its ranks can be estimated at more than 30,000 men—perhaps an insignificant number compared with that of modern governments but gigantic when set against the total of only a few hundred a century earlier. The problem, however, lay not in numbers but in the assumption, held throughout both bureaucracy and army, that a position of power ex officio entitled the holder to a rake-off of some sort, to be extracted both from the citizenry with whom he came in contact and from fellow members of the service in ranks below his own. This ethos was not new; but during the principate it had been restrained by higher officers and officials, who operated according to a different, essentially aristocratic, code expressed in patron-dependent relations and mutuality. Its currency was not money but favours and services. Such a code was swept away by the rapid increase in the size of government in the later 3rd century and the rise to high civil and military posts by men recruited from the ranks rather than from the upper classes. As they had bought their own promotions or appointments, so they expected to recoup their expenses (and more besides) by such means as selling exemptions and extortion. The more intrusive and demanding the military tax collection or the state's control of the rosters of city senates, the more profit there was for a pervasively corrupt administration. Persons close to the emperor could, for a price, generally screen him from knowledge of what was going on. Constantine, for example, complained quite in vain—and the complaint was endlessly repeated by his successors—that the city senates were being "emptied of persons obligated to them by birth, who yet are asking for a government post by petition to the emperor, running off to the legions or various civil offices." Such posts could easily be bought. A great deal of imperial planning was thus vitiated by sale. Many of the profiteers started life in the urban upper classes, but, as nouveaux riches, they joined the older landed nobility after a term in the emperor's service.

Roger-Viollet



Mosaic of a fortified Roman villa with lookout towers and an orchard, in the Le Bardo National Museum, Bārdaw, Tunisia.

Consolidation of landownership

In a few areas where measurement is possible, one can see that a process of consolidation of landownership had been going on for a long time, bringing the rural population increasingly into dependence on the larger property holders. Diocletian's new system of property assessment accelerated this process; it was more thorough and thus exposed the poor and ignorant to exploitation by local officialdom. In response, they sought the protection of some influential man to ward off unfair assessments, selling their land to him and becoming his tenants. In areas disturbed by lawlessness, a large landowner offered them safety as well. The strength of rural magnates in their formidable, even fortified, dwellings, with a dependent peasantry of 100 or even 1,000 around them made much trouble for tax collectors, and landowners thus became the target of many laws. Consolidation of ownership, however, was not apparent in northern Africa, and the reverse process has been established for a carefully researched area of Syria.

Regional differences cannot be disregarded. They were responsible for guiding the development of the later empire along quite varied paths. The archaeological data, which reflect these developments most clearly, register such changes as the degree of wealth in public buildings and the use and elegance of carved sarcophagi or of mosaics in private houses. Broadly speaking, a decline is noticeable throughout the European provinces; it tends to affect the cities earlier than the rural areas and is detectable sometimes by 350, generally by 375. In the Danube provinces, the evidence fits neatly with political history following the Battle of Adrianople in 378, after which their condition was continually disturbed by the Visigothic immigrants. There is, however, no such obvious explanation for areas such as Spain or central Gaul. Italy of the 3rd and 4th centuries was not perceptibly worse off than before, though wealth in the Po region was more concentrated in the cities north of the river. Northern Africa seems to have maintained nearly the same level of prosperity as in earlier centuries, if proper weight is given to ecclesiastical building after Constantine. For Egypt, no clear picture emerges; but all the other Eastern provinces enjoyed in the later empire the same level of economic well-being as before or a still higher one, with more disposable wealth and an increasing population. These conditions continued into the 5th century.

The vast differences between the European and the Eastern provinces are best explained by the shifting focus of imperial energies. It can be traced in the locus of heaviest

military recruitment, in the lower Danube, as the 3rd century progressed; in the consequent concentration of military expenditure there; and in the siting of the emperors' residence as it was moved from Rome to Milan in the 260s, then to the lower Danube later in the 3rd century (where much fighting occurred), and subsequently to Nicomedia (Diocletian's capital). None of the Tetrarchs chose Rome—its days as the imperial centre were over—and when, from among various Eastern cities he considered, Constantine decided on Byzantium as his permanent residence, he simply made permanent a very long-term development. Meanwhile the Rhine frontier and the upper Danube were repeatedly overrun. As can be inferred from the signs of fortification in Pannonia, Gaul, Britain, and Spain, internal policing was neglected. Commercial intercourse, which had been the key to raising the economy and the level of urbanization, became less safe and easy. Villas turned into self-sufficient villages, and the smaller towns also reverted to villages. Only the larger towns, such as Bordeaux, Arles, or Cartagena, maintained their vitality.

Although there was considerable inflation (culminating under Theodosius), in spite of a deflationary fiscal policy, commercial transactions ignored barter and were based instead on currency throughout the empire at the end of the century. The economy was partially under state direction, which was applied to agriculture through bias toward the settler system on imperial estates and to industry through the requisitioning of corporations (artisans, merchants, carriers) and the creation of state workshops (especially for manufacturing military goods). Opinions differ on the intensity of trade, but there was certainly clear progress in comparison to the 3rd century.

The remnants of pagan culture. The spread of Christianity in no way harmed the flourishing of pagan literature. Instruction in the universities (Rome, Milan, Carthage, Bordeaux, Athens, Constantinople, Antioch, Alexandria) was still based on rhetoric, and literature received the support of senatorial circles, especially in Rome (for example those of the Symmachi and the Nicomachi Flaviani). Latin literature was represented by Symmachus and the poet Ausonius. The last great historian of Rome was Ammianus Marcellinus, a Greek who wrote in Latin for the Roman aristocracy; of his *Res gestae,* the most completely preserved part describes the period from 353 to 378. The works of Sextus Aurelius Victor and Eutropius, who ably abridged earlier historical works, are fairly accurate and more reliable than the *Scriptores historiae Augustae,* a collection of imperial biographies of unequal value, undoubtedly composed under Theodosius but for an unknown purpose. Erudition was greatly prized in aristocratic circles, which, enamoured of the past, studied and commented on the classic authors (Virgil) or the ancestral rites (the *Saturnalia* of Macrobius). Greek literature is represented by the works of philosophers or sophists: Themistius, a political theoretician who advocated absolutism; Himerius of Prusias; and above all Libanius of Antioch, whose correspondence and political discourses from the Theodosian period bear witness to his perspicacity and, often, to his courage.

The Christian church. In the last decade of the 4th century the harsh laws against the perpetuation of the old pieties promulgated by Theodosius gave impetus and justification to waves of icon and temple destruction, especially in the East. It is, nonetheless, likely that a majority of the population was still non-Christian in 400, although less so in the cities and in the East and more so in rural and mountainous areas and the West. Efforts by the church to reach them were intermittent and lacked energy. Bishops generally expected rural magnates to do their job for them; and the church leadership was, in any case, of a social class that viewed the peasantry from a great distance and wanted to keep it that way. Except by such unusual figures as Martin of Tours or Marcellus of Apamea, little effort was made to convert people who were hard to reach. As always in antiquity, it was in the cities where changes occurred—with the exception of monasticism.

Monasticism

Only in the reign of Constantine, and about simultaneously in Egypt and Palestine, had monasticism, a religious movement whose followers lived as hermits and pursued a life of extreme asceticism, become more than the little-regarded choice of rare zealots. Near Gaza and in the desert along the eastern side of Jerusalem a number of tiny clusters of cells had been made from caves and taken as residence by ascetics, from whose fame and example that way of life later spread to many other corners of the Levant. The bishop of Jerusalem, Cyril, by mid-century could speak of "regiments of monks." But it was in the desert on both sides of the Nile that similar ascetic experiments of much greater importance were made, by the hero of the movement, St. Anthony, and others. True monasticism, tempered only by weekly communal worship and organizing, established itself on Anthony's model and under his inspiration in the first decade of the 4th century; it took root above all in the desert of Scete just west of the base of the Nile delta. Coenobitism, joint life in enclosed communities, was the model preferred by St. Pachomius around 330, vigorously directed and diffused by him until mid-century, when both he and Anthony died. Basil of Caesarea was to establish monastic communities in Cappadocia under the influence of what he saw in Egypt on his visit there in the year of Anthony's death; and Athanasius was shortly to write a biography of that saint of enormous influence and to carry word of his life to Italy and Gaul during his own exile there. The biography was soon translated into Latin and inspired a scattering of experiments in asceticism or coenobitism in the West—in Vercellae in Italy, for example, by 330, and at Tours in the 370s under Martin's direction. Tours became the first monastery in the West comparable to those in the East, but development subsequently was slow compared to the 10,000 or more monasteries founded in Egypt by AD 400.

The most distinct and well-reported phenomenon during the century after the conversion of Constantine was the continued religious rioting and harrying in the cities, both in all the major ones and in dozens of minor ones. The death toll exceeded the toll among Christians at the hands of pagans in earlier persecutions. It was rarely of Jews or pagans at the hands of Christians or of Christians at the hands of pagans, but ordinarily of Christians at each others' hands in the course of sectarian strife. For a time, no one sect enjoyed a majority among Homoousians, Arians, Donatists, Meletians, and many others. Bishoprics were fiercely contested and appeals often made to armed coercion. The emperors had assumed the right to interfere and often did so; but under Theodosius, Pope Damasus and St. Ambrose reacted: the state was to restrict itself to furnishing the "secular arm," while the church, in the name of evangelical ethics, claimed the right to judge the emperors, a policy that had grave implications for the future. The "caesaro-papism" of Constantius later gained adherents under the Byzantine emperors. In the meantime, the Goths had been converted to Arianism by Ulfilas during the period of Constantius and Valens, thus presaging conflicts that were to come after the great invasions. Orthodox missionaries had converted Osroëne, Armenia, and even some countries on the Red Sea.

Triumph of orthodoxy

The Christian literature of the 4th century is remarkable. Its first representative is Eusebius of Caesarea, a friend and panegyrist of Constantine and a church historian whose creation of a "political theology" sealed the union between the Christian emperor and the church. St. Athanasius wrote apologetic works and a life of St. Anthony. Also prominent were the great Cappadocians: St. Basil of Caesarea, St. Gregory of Nazianzus, St. Gregory of Nyssa, and St. John Chrysostom of Antioch, the greatest preacher of his time. The Westerners, too, had great scholars and brilliant writers: St. Hilary of Poitiers, enemy of the Arians and of Constantius; St. Ambrose, administrator and pastor, whose excessive authority was imposed on Gratian and even on Theodosius; and St. Jerome, a desert monk and confessor of upper-class Roman ladies, a formidable polemicist who knew Greek and Hebrew and made the first faithful translation of the Old and New Testaments (the Vulgate) as well as of a chronicle of world history, which was a translation and continuation of the work of Eusebius. Finally, St. Augustine, the bishop of Hippo, was a great pastor, a vigorous controversialist, a sensitive and passionate writer (the *Confessions*), and the powerful

theologian of *The City of God.* The century that developed these great minds cannot be considered decadent.

The eclipse of the Roman Empire in the West (*c.* 395–500) and the German migrations. *Invasions in the early 5th century.* After the death of Theodosius the Western empire was governed by young Honorius. Stilicho, an experienced statesman and general, was charged with assisting him and maintaining unity with the East, which had been entrusted to Arcadius. The Eastern leaders soon rejected Stilicho's tutelage. An antibarbarian reaction had developed in Constantinople, which impeded the objectives of the half-Vandal Stilicho. He wanted to intervene on several occasions in the internal affairs at Constantinople but was prevented from doing so by a threat from the Visigoth chieftain Alaric, whom he checked at Pollentia in 402, then by the Ostrogoth Radagaisus' raid in 406, and finally by the great invasion of the Gauls in 407. The following year he hoped to restore unity by installing a new emperor in Constantinople, Theodosius II, the son of Arcadius, who had died prematurely; but he succumbed to a political and military plot in August 408. The division of the two *partes imperii* was now a permanent one.

Honorius, seated in Ravenna, a city easier to defend than Milan, had only incompetent courtiers surrounding him, themselves animated by a violent hatred of the barbarians.

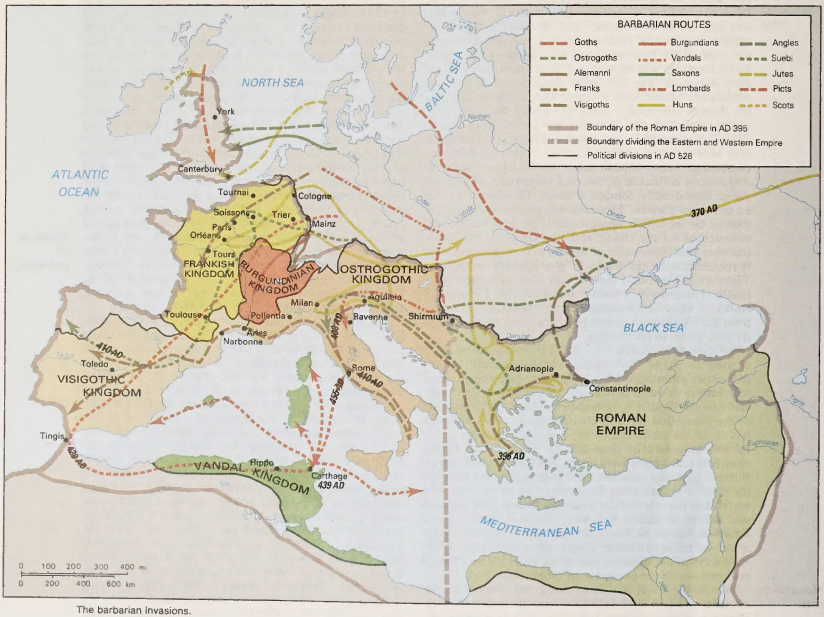
Sack of Rome by the Visigoths

Alaric soon reappeared, at the head of his Visigoths, demanding land and money. Tired of the Romans' double-dealing, he descended on Rome itself. The city was taken and pillaged for three days, thus putting an end to an era of Western history (August 410). An Arian, Alaric spared the churches. He died shortly thereafter in the south; his successor, Athaulf, left the peninsula to march against the Gauls.

Fleeing from the terrifying advance of the Huns, on Dec. 31, 406, the Vandals, Suebi, and Alani, immediately followed by the Burgundians and bands of Alemanni, crossed the frozen Rhine and swept through Gaul, effortlessly throwing back the federated Franks and Alemanni from the frontiers. Between 409 and 415 a great many of these barbarians arrived in Spain and settled in Lusitania (Suebi) and in Baetica (Vandals, whence the name Andalusia). As soon as Gaul had become slightly more peaceful, Athaulf's Visigoths arrived, establishing themselves in Narbonensis and Aquitania. After recognizing them as "federates," Honorius asked them to go to Spain to fight the Vandals. Meanwhile, the Roman general Constantius eliminated several usurpers in Gaul, confined the Goths in Aquitania, and reorganized the administration (the Gallic assembly of 418). But he was unable to expel the Franks, the Alemanni, and the Burgundians, who had occupied the northern part of the country, nor to eliminate the brigandage of the Bagaudae. He was associated with the empire and was proclaimed Augustus in 421, but he died shortly afterward. His son, Valentinian III, succeeded Honorius in 423 and reigned until 455.

The beginning of Germanic hegemony in the West. During the first half of the 5th century the barbarians gradually installed themselves, in spite of the efforts of the Roman general Aetius at the head of a small army of mercenaries and of Huns. Aetius took back Arles and Narbonne from the Visigoths in 436, either pushed back the Salian and Ripuarian Franks beyond the Rhine or incorporated them as federates, settled the defeated Burgundians in Sapaudia (Savoy), and established the Alani in Orléans. The other provinces were lost: Britain, having been abandoned in 407 and already invaded by the Picts and Scots, fell to the Angles, Saxons, and Jutes; a great Suebi kingdom, officially federated but in fact independent, was organized

Loss of provinces



The barbarian invasions.

in Spain after the departure of the Vandals, and it allied itself to the Visigoths of Theodoric I, who were settled in the country around the Garonne.

In 428 the Vandal Gaiseric led his people (80,000 persons, including 15,000 warriors) to Africa. St. Augustine died in 430 in besieged Hippo, Carthage fell in 435, and in 442 a treaty gave Gaiseric the rich provinces of Byzacena and Numidia. From there he was able to starve Rome, threaten Sicily, and close off the western basin of the Mediterranean to the Byzantines.

Shortly afterward, in 450, Attila's Huns invaded the West—first Gaul, where, after having been kept out of Paris, they were defeated by Aetius on the Campus Mauriacus (near Troyes), then Italy, which they evacuated soon after having received tribute from the pope, St. Leo. Attila died shortly afterward; and this invasion, which indeed left more legendary memories than actual ruins, had shown that a solidarity had been created between the Gallo-Romans and their barbarian occupiers, for the Franks, the Alemanni, and even Theodoric's Visigoths had come to Aetius' aid.

After the death of Aetius, in 454, and of Valentinian III, in 455, the West became the stake in the intrigues of the German chiefs Ricimer, Orestes, and Odoacer, who maintained real power through puppet emperors. In 457–461 the energetic Majorian reestablished imperial authority in southern Gaul until he was defeated by Gaiseric and assassinated shortly afterward. Finally, in 476, Odoacer deposed the last emperor, Romulus Augustulus, had himself proclaimed king in the barbaric fashion, and governed Italy with moderation, being de jure under the emperor of the East. The end of the Roman Empire of the West passed almost unperceived.

The end of the Roman Empire of the West

Barbarian kingdoms. Several barbarian kingdoms were then set up: in Africa, Gaiseric's kingdom of the Vandals; in Spain and in Gaul as far as the Loire, the Visigothic kingdom; and farther to the north, the kingdoms of the Salian Franks and the Alemanni. The barbarians were everywhere a small minority. They established themselves on the great estates and divided the land to the benefit of the federates without doing much harm to the lower classes or disturbing the economy. The old inhabitants lived under Roman law, while the barbarians kept their own "personality of laws," of which the best-known is the judicial composition, the Wergild. Romans and barbarians coexisted but uneasily. Among the obstacles to reconciliation were differences in mores; social and political institutions (personal monarchies, fidelity of man to man); language, although Latin was still used in administration; and above all religion: the Arianism of the barbarians permitted the Roman Catholic bishops to retain their hold over their flocks. The only persecution, however, was under the Vandals, whose domination was the harshest.

Two great kingdoms marked the end of the 5th century. In Gaul, Clovis, the king of the Salian Franks (reigned 481/482–511), expelled Syagrius, the last Roman, from Soissons, took Alsace and the Palatinate from the Alemanni (496), and killed Alaric II, king of the Visigoths, at Vouillé (507). His conversion to Catholicism assured him the support of the bishops, and Frankish domination was established in Gaul. At the same time, Theodoric, king of the Ostrogoths, reigned in Italy. He had been charged by the emperor Zeno to take back Italy from Odoacer in 488, and in 494 he had himself proclaimed king at Ravenna. His Goths, few in number, were established in the north; elsewhere he preserved the old imperial administration, with senators as prefects. Externally, he kept Clovis from reaching the Mediterranean and extended his state up to the valley of the Rhône. Theodoric died in 526. Ten years later Justinian charged his general Belisarius with the reconquest of Italy, a costly, devastating, and temporary operation that lasted from 535 to 540.

Analysis of the decline and fall. The causes of the fall of the empire have been sought in a great many directions and with a great deal of interest, even urgency, among historians of the West; for it has been natural for them to see or seek parallels between Rome's fate and that of their own times. In any choice of explanations there is likely to be a hidden sense of priorities determining the definition of "civilization," or specifically the civilization of "Rome" or of "the classical world." If, for example, classical civilization is identified with the literature of the ancients at what one conceives to be its best, then "the end" of this civilization has to be set at some point of decline and explanations for its coming to be sought in the preexisting conditions; or if not on literature but on political domination, then some other point in time must be chosen and explained in terms of what seems to have led up to it. There have been endless variations on this search, and there will continue to be more, no doubt, since it is agreed that literature did, in fact, diminish in quality, as did jurisprudence, although at a different date, and oratory, and vigorous political debate in the capital, and powerfully innovative philosophy, and sculpture, and civic patriotism, and the willingness to die for one's country. "Civilization" turns out to be not one single entity but a web of many strands, each of its own length.

Perhaps the view attracting the most adherents, however, has focused on the ability of the empire to maintain its political and military integrity—that being the strand apparently most central and significant—and the juncture at which that ability is most dramatically challenged and found wanting—the period of "the barbarian invasions," meaning 407 and roughly the ensuing decade. If this juncture in turn is examined and the antecedents of the empire's weakness sought in internal developments, they can only be found in the government. Belief in and obedience to the monarch was not lacking, nor military technology at least matching that of the invaders, nor a population large enough to field a large force, nor the force itself (on paper, at least), nor the economic potential adequate to the arming of it. Particular defeats described by contemporaries in reasonable detail are almost uniformly attributable to the rottenness of government as described above, rendering soldiers undisciplined, untrained, frequently on indefinite leave, and without good morale or proper equipment. Soldiers were unpaid because of various abuses in the collection and delivery of supplies and money from taxpayers, and they were distracted from their proper duties by their own and their officers' extortionate habits in contact with their civilian hosts. For the same basic reason—that is, abuse of power wielded through service in the army or bureaucracy—the administration of the cities no longer enjoyed the efforts of the urban elites, who by 407 had long since fled from active service to some exempt government post or title. For the same reason, finally, corrective measures needed against these systemic weaknesses could not be developed by enlightened men at the centre because they were screened from the truth of things, were at the mercy of incompetent or venal agents, or were unable to maintain themselves in power against the plotters around them. The details of all these charges that can be made against late Roman government are writ large in the great collection of imperial edicts published in 438, the Theodosian Code, as well as in the works of roughly contemporary writers from East and West, such as Synesius, Augustine, Libanius, Themistius, Chrysostom, Symmachus, Bishop Maximus of Turin, and, above all, Ammianus Marcellinus. An empire that could not deliver to a point of need all the defensive force it still possessed could not well stand against the enemy outside.

(P.P./R.MacM.)

BIBLIOGRAPHY

General works on ancient Greek and Roman civilizations. A wealth of information on ancient Greek and Roman civilizations is provided by the volumes in *The Cambridge Ancient History* (1923–), some in newer 2nd and 3rd editions; by N.G.L. HAMMOND and H.H. SCULLARD (eds.), *The Oxford Classical Dictionary*, 2nd ed. (1970, reprinted 1984); and by JOHN BOARDMAN, JASPER GRIFFIN, and OSWYN MURRAY (eds.), *The Oxford History of the Classical World* (1986). MICHAEL GRANT and RACHEL KITZINGER (eds.), *Civilization of the Ancient Mediterranean: Greece and Rome*, 3 vol. (1988), discusses the geography, inhabitants, arts, language, religion, politics, technology, and economy of the area from the early 1st millennium BC to the late 5th century AD. Broad coverage of the physical and cultural settings and of archaeological discoveries is also provided by PETER LEVI, *Atlas of the Greek World* (1980); TIM CORNELL and JOHN MATTHEWS, *Atlas of the Roman World*

(1982); and NICHOLAS G.L. HAMMOND (ed.), *Atlas of the Greek and Roman World in Antiquity* (1981). Overviews of the histories of each civilization include, on Greece, NICHOLAS G.L. HAMMOND, *A History of Greece to 322 B.C.*, 3rd ed. (1986); J.B. BURY and RUSSELL MEIGGS, *A History of Greece to the Death of Alexander*, 4th ed. (1975), revised to take account of new evidence; and ROGER LING, *The Greek World* (1976, reissued 1990); and, on Rome, M. CARY and H.H. SCULLARD, *A History of Rome Down to the Reign of Constantine*, 3rd ed. (1975); and MICHAEL VICKERS, *The Roman World* (1977, reissued 1989). Many ancient historical sources are available in *The Loeb Classical Library* series, with original text and parallel English translation; in the series *Translated Documents of Greece and Rome*; and in NAPHTALI LEWIS and MEYER REINHOLD (eds.), *Roman Civilization: Selected Readings*, 3rd ed., 2 vol. (1990).

(E.D.T.V./S.H./J.Fe./N.T.deG./G.E.Fo./R.P.Sa./R.MacM.)

Aegean civilizations. General overviews include EMILY VERMEULE, *Greece in the Bronze Age* (1964, reissued 1974), the standard work; HANS-GÜNTHER BUCHHOLZ and VASSOS KARAGEORGHIS, *Prehistoric Greece and Cyprus: An Archaeological Handbook* (1973; originally published in German, 1971); SPYRIDON MARINATOS and MAX HIRMER, *Kreta, Thera, und das mykenische Hellas*, 3rd ed. (1976), also available in an English translation of an earlier edition, *Crete and Mycenae* (1960); WILLIAM TAYLOUR, *The Mycenaeans*, rev. ed. (1983); N.K. SANDARS, *The Sea Peoples: Warriors of the Ancient Mediterranean, 1250–1150 B.C.*, rev. ed. (1985); and PETER WARREN, *The Aegean Civilizations* (1975, reissued 1989). Ancient Crete is discussed in ARTHUR EVANS, *The Palace of Minos*, 4 vol. (1921–35, reissued 1964), still basic; ARTHUR EVANS, MARK CAMERON, and SINCLAIR HOOD, *Knossos Fresco Atlas* (1967); SINCLAIR HOOD, *The Minoans* (1971); and J. WALTER GRAHAM, *The Palaces of Crete*, rev. ed. (1987). JOHN BOARDMAN, *The Cretan Collection in Oxford* (1961), describes Cretan antiquities in the Ashmolean Museum. The Cycladic civilizations are examined by JÜRGEN THIMME (ed.). *Art and Culture of the Cyclades* (also published as *Art and Culture of the Cyclades in the Third Millennium B.C.*, 1977; originally published in German, 1976), an extensive exhibition catalog; CHRISTOS G. DOUMAS, *Thera, Pompeii of the Ancient Aegean* (1983); and R.L.N. BARBER, *The Cyclades in the Bronze Age* (1987). Ancient civilization on the Greek mainland is the focus of GEORGE E. MYLONAS, *Mycenae and the Mycenaean Age* (1966), for the Late Bronze Age and Mycenaean Shaft Grave Circle B; EMILY VERMEULE, *The Art of the Shaft Graves of Mycenae* (1975); and I.T. HOOKER, *Mycenaean Greece* (1976). Religion and religious sites are discussed in MARTIN P. NILSSON, *The Minoan-Mycenaean Religion and Its Survival in Greek Religion*, 2nd rev. ed. (1950, reprinted 1971), still the standard work; and BOGDAN RUTKOWSKI, *The Cult Places of the Aegean* (1986). Information on ancient pottery and seals may be found in ARNE FURUMARK, *The Mycenaean Pottery: Analysis and Classification* (1941, reissued 1972); FRIEDRICH MATZ *et al.*, *Corpus der Minoischen und Mykenischen Siegel* (1964–); and JOHN BOARDMAN, *Greek Gems and Finger Rings* (1970). MICHAEL VENTRIS and JOHN CHADWICK, *Documents in Mycenaean Greek*, 2nd ed. (1973), is essential for information on the writing and decipherment of Linear B, including transcriptions, translations, and commentary on selected tablets. For the Aegean Bronze Age as a background to Homer, see the series *Archaeologia Homerica: die Denkmäler und das frühgriechische Epos*, ed. by FRIEDRICH MATZ and HANS-GÜNTER BUCHHOLZ (1967–). SINCLAIR HOOD, *The Arts in Prehistoric Greece* (1978, reissued 1988), is the best standard work on this subject. WILLIAM S. SMITH, *Interconnections in the Ancient Near-East: A Study of the Relationships Between the Arts of Egypt, the Aegean, and Western Asia* (1965).

(M.S.F.H./E.D.T.V.)

Ancient Greek civilization. *The early Archaic period:* JOHN BOARDMAN, *The Greeks Overseas: Their Early Colonies and Trade*, rev. and enlarged ed. (1980), is a well-illustrated and fully documented account of Greek colonization. Commercial factors are stressed by MARTIN FREDERIKSEN, *Campania*, ed. by NICHOLAS PURCELL (1984). PETER GARNSEY, in *Famine and Food Supply in the Graeco-Roman World* (1988), is skeptical of the "land-hunger" explanation. See also ROBERT SALLARES, *The Ecology of the Ancient Greek World* (1991), a discussion of problems of demography and food supply. The importance of rural sanctuaries to the growth of the polis is argued in FRANÇOIS DE POLIGNAC, *La Naissance de la cité grecque* (1984); and ROBIN OSBORNE, *Classical Landscape with Figures* (1987), which also discusses the ways in which the Greek countryside was exploited in different regions. "Monumentalization" is stressed in a good general account of the period, ANTHONY SNODGRASS, *Archaic Greece: The Age of Experiment* (1980). IAN MORRIS, *Burial and Ancient Society* (1987), discusses burial and the Greek polis. MARTIN BERNAL, *Black Athena: The Afroasiatic Roots of Classical Civilization* (1987–), is a controversial work about Phoenician influence on Greece.

Explorations of phratry and *genos* are found in DENIS ROUSSEL, *Tribu et cité* (1976); and FÉLIX BOURRIOT, *Recherches sur la nature du Genos*, 2 vol. (1976). K.J. DOVER, *Greek Homosexuality*, new ed. (1989), is a good treatment. The significance of the symposium is argued for by OSWYN MURRAY, "The Symposion as Social Organisation," in ROBIN HÄGG (ed.), *The Greek Renaissance of the Eighth Century B.C.: Tradition and Innovation* (1983), pp. 195–199. Murray's *Early Greece* (1980) is a readable general history of the period stressing the symposium at a number of points. GABRIEL HERMAN, *Ritualised Friendship and the Greek City* (1987), examines *xenia*. A discussion of *proxenia* may be found in L.H. JEFFERY, *Archaic Greece: The City-States, c. 700–500 B.C.* (1976), an elegant general history arranged regionally. ROBERT DREWS, *Basileus: The Evidence for Kingship in Geometric Greece* (1985), argues that the early Greek "kings" (*basileis*) were really just hereditary aristocrats; but see PIERRE CARLIER, *La Royauté en Grèce avant Alexandre* (1984), for a differing view.

Useful monographs on individual poleis important in the Archaic period include J.B. SALMON, *Wealthy Corinth: A History of the City to 338 BC* (1984); and THOMAS J. FIGUEIRA, *Aegina: Society and Politics* (1981). But see M.I. FINLEY, *Politics in the Ancient World* (1983), which warns against the dangers of writing histories of particular poleis.

The later Archaic periods: The classic exposition of the "hoplite theory" of tyranny is A. ANDREWES, *The Greek Tyrants* (1956, reissued 1974); it is refined by PAUL CARTLEDGE, "Hoplites and Heroes: Sparta's Contribution to the Technique of Ancient Warfare," *Journal of Hellenic Studies*, 97:11–27 (1977). But G.L. CAWKWELL, *Philip of Macedon* (1978), ch. 10, discusses hoplite fighting as a more individual affair than is sometimes allowed. The ideological implications of hoplite fighting are treated by W.R. CONNOR, "Early Greek Land Warfare as Symbolic Expression," *Past & Present*, 119:3–29 (May 1988); and PIERRE VIDAL-NAQUET, *The Black Hunter: Forms of Thought and Forms of Society in the Greek World* (1986; originally published in French, 1981).

W.G. FORREST, *A History of Sparta, 950–192 B.C.*, 2nd ed. (1980), is a provocative brief work. PAUL CARTLEDGE, *Sparta and Lakonia: A Regional History, 1300–362 BC* (1979), is also useful. "Laconism" is explored in ELIZABETH RAWSON, *The Spartan Tradition in European Thought* (1969). R.E. WYCHERLEY, *The Stones of Athens* (1978), discusses Athens' natural advantages. Attic produce and exports are studied by SIGNE ISAGER and MOGENS HERMAN HANSEN, *Aspects of Athenian Society in the Fourth Century B.C.* (1975). RUSSELL MEIGGS, *Trees and Timber in the Ancient Mediterranean World* (1982); and ROBERT GARLAND, *The Piraeus: From the Fifth to the First Century B.C.* (1987), present naval aspects. Discussion of every aspect of early Athenian political history is contained in P.J. RHODES, *Commentary on the Aristotelian Athenaion Politeia* (1981). ROBERT PARKER, *Miasma: Pollution and Purification in Early Greek Religion* (1983), covers Cylon and the religious taint often incurred through some wrongful act or neglect of ritual obligation. M.I. FINLEY, *Ancient Slavery and Modern Ideology* (1980), investigates the link between Solon and slavery.

A good account of Peisistratid building policy in its competitive aspect is found in T. LESLIE SHEAR, JR., "Tyrants and Buildings in Archaic Athens," in *Athens Comes of Age: From Solon to Salamis* (1978), pp. 1–15. Peisistratid artistic propaganda is summarized in JOHN BOARDMAN, "Archaic Greek Society: Material Culture," ch. 7c in JOHN BOARDMAN *et al.* (eds.), *The Cambridge Ancient History*, vol. 4, 2nd ed. (1988), pp. 414–430. R.M. COOK, "Pots and Peisistratan Propaganda," *Journal of Hellenic Studies*, 107:167–169 (1987), presents another view. JOHN S. TRAILL, *Demos and Trittys: Epigraphical and Topographical Studies in the Organization of Attica* (1986), a specialist account of the deme system based on inscriptions, supplements his *The Political Organization of Attica: A Study of the Demes, Trittyes, and Phylai, and Their Representation in the Athenian Council* (1975). DAVID WHITEHEAD, *The Demes of Attica, 508/7–ca. 250 B.C.: A Political and Social Study* (1986), is magisterial and reliable. ROBIN OSBORNE, *Demos: The Discovery of Classical Attika* (1985), is a speculative essay with some interesting suggestions not fully worked out. EMILY KEARNS, "Change and Continuity in Religious Structures after Kleisthenes," in PAUL CARTLEDGE and F.D. HARVEY (eds.), *Crux* (1985), pp. 189–207, is valuable on the religious aspects of Cleisthenes' reforms. The modern reconstructed Greek trireme is the subject of J.S. MORRISON and J.F. COATES, *The Athenian Trireme* (1986).

Early Greek philosophy is dealt with generally in EDWARD HUSSEY, *The Presocratics* (1972); and JONATHAN BARNES, *The Presocratic Philosophers*, rev. ed. (1982), a more difficult work; and on Pherecydes in particular in M.L. WEST, *Early Greek Philosophy and the Orient* (1971). More works on Greek philosophy and philosophers can be found in the bibliography of the article THE HISTORY OF WESTERN PHILOSOPHY. ROSALIND

THOMAS, *Oral Tradition and Written Record in Classical Athens* (1989), supersedes all previous studies on literacy.

Classical Greek civilization. Information on the Persian empire and the Ionian revolt can be found in J.M. COOK, *The Persian Empire* (1983); and a postscript by DAVID M. LEWIS in the book by A.R. BURN, *Persia and the Greeks*, 2nd ed. (1984). Lewis' *Sparta and Persia* (1977), analyzes Persian administration. SIMON HORNBLOWER, *Mausolus* (1982), treats satrapally controlled Anatolia.

General histories of the period 479 BC to Alexander the Great include J.K. DAVIES, *Democracy and Classical Greece* (1978); and SIMON HORNBLOWER, *The Greek World, 479–323 BC*, rev. ed. (1991). RUSSELL MEIGGS, *The Athenian Empire* (1972), offers a full scholarly history and analysis. A very different view, in particular denying that there was a mid-century qualitative change in the character of the Athenian empire, is presented by M.I. FINLEY, "The Fifth-Century Athenian Empire: A Balance-Sheet," in PETER GARNSEY and C.R. WHITTAKER (eds.), *Imperialism in the Ancient World* (1978), pp. 103–126. More succinct than Meiggs is the very brief work by P.J. RHODES, *The Athenian Empire* (1985). G.W.M. DE STE. CROIX, *The Class Struggle in the Ancient Greek World: From the Archaic Age to the Arab Conquests* (1981), contains much relevant material. Athenian foreign policy in the period is explored by E. BADIAN, "The Peace of Callias," *Journal of Hellenic Studies*, 107:1–39 (1987); and DAVID M. LEWIS, "The Origins of the First Peloponnesian War," in GORDON SPENCER SHRIMPTON and DAVID JOSEPH MCCARGAR (eds.), *Classical Contributions* (1981), pp. 71–78. Works on Thucydides and his *History* include W. ROBERT CONNOR, *Thucydides* (1984); SIMON HORNBLOWER, *Thucydides* (1987); COLIN MACLEOD, *Collected Essays* (1981); NICOLE LORAUX, *The Invention of Athens: The Funeral Oration in the Classical City* (1986); and A.W. GOMME, A. ANDREWES, and K.J. DOVER, *A Historical Commentary on Thucydides*, 5 vol. (1945–81). This last work assumes a good knowledge of Greek, but SIMON HORNBLOWER, *A Commentary on Thucydides* (1991–), translates all Greek commented on.

P.J. RHODES, *The Athenian Boule* (1972, reissued 1985), discusses an important aspect of internal Athenian history in the 5th century. A full account of constitutional developments is given in MARTIN OSTWALD, *From Popular Sovereignty to the Sovereignty of Law* (1986). J.K. DAVIES, *Athenian Propertied Families, 600–300 B.C.* (1971), is indispensable on individual politicians; it may be supplemented by Davies' *Wealth and the Power of Wealth in Classical Athens* (1981). A stimulating treatment of demagogues is given in W. ROBERT CONNOR, *The New Politicians of Fifth-century Athens* (1971); but WESLEY E. THOMPSON, "Athenian Leadership: Expertise or Charisma?" in GORDON SPENCER SHRIMPTON and DAVID JOSEPH MCCARGAR (eds.), *Classical Contributions* (1981), pp. 153–159, argues against all attempts to impute greater professionalism to them. Studies on oligarchic sympathizers include L.B. CARTER, *The Quiet Athenian* (1986); GLENN RICHARD BUGH, *The Horsemen of Athens* (1988); and ANDREW LINTOTT, *Violence, Civil Strife, and Revolution in the Classical City, 750–330 B.C.* (1981).

G.E.M. DE STE. CROIX, *The Origins of the Peloponnesian War* (1972), argues a controversial thesis, but is excellent on Sparta. A contribution on this topic of the first importance is E. BADIAN, "Thucydides and the Outbreak of the Peloponnesian War," in JUNE W. ALLISON (ed.), *Conflict, Antithesis, and the Ancient Historian* (1990), pp. 46–91, showing that Thucydides' presentation has suspiciously pro-Athenian features. Athenian strategy in the war itself is discussed in J.B. SALMON, *Wealthy Corinth: A History of the City to 338 BC* (1984); G.L. CAWKWELL, "Thucydides' Judgment of Periclean Strategy," *Yale Classical Studies*, 24:53–70 (1975); and A.J. HOLLADAY, "Athenian Strategy in the Archidamian War," *Historia*, 27(3):399–426 (1978). Spartan strategy is examined by P.A. BRUNT, "Spartan Policy and Strategy in the Archidamian War," *Phoenix*, 19 (4):255–280 (1965). The link between speculative thinking and democracy is argued for by G.E.R. LLOYD, *Magic, Reason, and Experience: Studies in the Origin and Development of Greek Science* (1979); it is qualified in the epilogue to HUGH LLOYD-JONES, *The Justice of Zeus*, 2nd ed. (1983). See also CYNTHIA FARRAR, *The Origins of Democratic Thinking: The Invention of Politics in Classical Athens* (1988).

The best handbook on Greek art is MARTIN ROBERTSON, *A History of Greek Art*, 2 vol. (1975). JOHN BOARDMAN, *Greek Art*, new rev. ed. (1985), is also worth consulting. Greek tragedy is assessed in BRIAN VICKERS, *Towards Greek Tragedy: Drama, Myth, Society* (1973); and SIMON GOLDHILL, *Reading Greek Tragedy* (1986), which attempts to put Greek tragedy in its polis framework, and "The Great Dionysia and Civic Ideology," *Journal of Hellenic Studies*, 107:58–76 (1987). The standard work on Attic dramatic festivals is ARTHUR PICKARD-CAMBRIDGE, *The Dramatic Festivals of Athens*, 2nd ed. rev. by JOHN GOULD and DAVID M. LEWIS (1988). Further works may be found in the bibliographies of the articles GREEK DRAMATISTS and GREEK LITERATURE. Fifth-century Athenian building is put in a political context in JOHANNES SIPKO BOERSMA, *Athenian Building Policy from 561/0 to 405/4 B.C.* (1970); and the brief and provocative work by RHYS CARPENTER, *Architects of the Parthenon* (1970).

A useful general book about Greek women in classical antiquity is SARAH B. POMEROY, *Goddesses, Whores, Wives, and Slaves* (1975). The position of Athenian women is discussed in the splendid essay by JOHN GOULD, "Law, Custom, and Myth: Aspects of the Social Position of Women in Classical Athens," *Journal of Hellenic Studies*, 100:38–59 (1980); and by DAVID M. SCHAPS, *Economic Rights of Women in Ancient Greece* (1979). A valuable discussion of female religious life in mostly male-dominated Attica is CHRISTIANE SOURVINOU-INWOOD, *Studies in Girls' Transitions: Aspects of the Arkteia and Age Representation in Attic Iconography* (1988), which discusses Brauron and the Artemis cult, celebrated there by women and girls. AVERIL CAMERON and AMÉLIE KUHRT (eds.), *Images of Women in Antiquity* (1983), is an interesting collection of papers.

Slavery is discussed by M.I. FINLEY (ed.), *Classical Slavery* (1987), and *Ancient Slavery and Modern Ideology* (1980); and by YVON GARLAN, *Slavery in Ancient Greece*, rev. and expanded ed. (1988; originally published in French, 1982). The best accounts of ancient Greek military technology are E.W. MARSDEN, *Greek and Roman Artillery*, 2 vol. (1969–71); A.W. LAWRENCE, *Greek Aims in Fortification* (1979); and JOSIAH OBER, *Fortress Attica: Defense of the Athenian Land Frontier, 404–322 B.C.* (1985).

The 4th century: M.I. FINLEY, *Ancient Sicily*, rev. ed. (1979), includes discussion of Dionysius I. The study of 4th-century Athenian democracy has been transformed by MOGENS HERMAN HANSEN, *The Athenian Ecclesia*, 2 vol. (1983–89), a collection of essays, and *The Athenian Assembly in the Age of Demosthenes* (1987; originally published in German, 1984). Additional works include R.K. SINCLAIR, *Democracy and Participation in Athens* (1988), an intelligent synoptic account; and MOGENS HERMAN HANSEN, *The Athenian Democracy in the Age of Demosthenes* (1991), the best comprehensive treatment. Lysander's role in the causes of the Corinthian war is admirably discussed by A. ANDREWES, "Spartan Imperialism?" in PETER GARNSEY and C.R. WHITTAKER (eds.), *Imperialism in the Ancient World* (1978), pp. 91–102. The whole period from 404 to 360 BC is discussed from a Spartan point of view by PAUL CARTLEDGE, *Agesilaos and the Crisis of Sparta* (1987). But G.L. CAWKWELL, "The Decline of Sparta," *Classical Quarterly*, new series, 33(2):385–400 (1983), presents a very different perspective denying that there was any serious shortage of manpower at Sparta in this period. The diplomacy of the period is presented in TIMOTHY T.B. RYDER, *Koine Eirene* (1965).

Contrasting verdicts on the Second Athenian Confederacy are presented by G.T. GRIFFITH, "Athens in the Fourth Century," in PETER GARNSEY and C.R. WHITTAKER (eds.), *Imperialism in the Ancient World* (1978), pp. 127–144; JACK CARGILL, *The Second Athenian League: Empire or Free Alliance?* (1981); and G.L. CAWKWELL, "Notes on the Failure of the Second Athenian Confederacy," *Journal of Hellenic Studies*, 101:40–55 (1981), and "Athenian Naval Power in the Fourth Century," *Classical Quarterly*, new series, 34(2):334–345 (1984). Mausolus' role in its breakup is addressed by SIMON HORNBLOWER, *Mausolus* (1982).

Studies on the Theban hegemony include J.A.O. LARSEN, *Greek Federal States* (1968); G.L. CAWKWELL, "Epaminondas and Thebes," *Classical Quarterly*, new series, 22:254–278 (1972); and JOHN BUCKLER, *The Theban Hegemony, 371–362 BC* (1980).

The rise of Macedon is portrayed in N.G.L. HAMMOND, G.T. GRIFFITH, and F.W. WALBANK, *A History of Macedonia*, 3 vol. (1972–88); and R. MALCOLM ERRINGTON, *A History of Macedonia* (1990; originally published in German, 1986). Philip and Alexander are placed into historical context by G.L. CAWKWELL, *Philip of Macedon* (1978); MILTIADES B. HATZOPOULOS and LOUISA D. LOUKOPOULOS (eds.), *Philip of Macedon* (1980), which includes good pictures of the Vergina tomb discoveries; A.B. BOSWORTH, *Conquest and Empire: The Reign of Alexander the Great* (1988), a masterly study, both scholarly and readable; and ROBIN LANE FOX, *Alexander the Great* (1973, reissued 1986), a lively work.

Fourth-century Greek emigration is discussed well by PAUL MCKECHNIE, *Outsiders in the Greek Cities in the Fourth Century B.C.* (1989). Greek attitudes to foreigners are explored by ARNOLDO MOMIGLIANO, *Alien Wisdom: The Limits of Hellenization* (1975). M.J. OSBORNE, *Naturalization in Athens*, 4 vol. in 3 (1981–83), treats grants of citizenship. "Euergetism" is the subject of PAUL VEYNE, *Le Pain et le cirque: Sociologie historique d'un pluralisme politique* (1976), also available in an abridged translation, *Bread and Circuses* (1990). WALTER BURKERT, *Greek Religion* (1985; originally published in German, 1977), is a full and brilliant study; other works may be

found in the bibliography of the article ANCIENT EUROPEAN RELIGIONS. (S.H.)

Hellenism. General historical works include F.W. WALBANK, *The Hellenistic World* (1981), a useful one-volume account by a leading authority; MICHAEL GRANT, *From Alexander to Cleopatra: The Hellenistic World* (1982), a well-written, reliable conspectus; PETER GREEN, *Alexander to Actium: The Historical Evolution of the Hellenistic Age* (1990), on every aspect of Hellenistic cultural and political history; W.W. TARN and G.T. GRIFFITH, *Hellenistic Civilisation*, 3rd ed. (1952, reissued 1975), a masterly, pioneering, and eminently readable study; JOHN FERGUSON, *The Heritage of Hellenism* (1973), thematic and well-illustrated; WILLIAM SCOTT FERGUSON, *Hellenistic Athens: An Historical Essay* (1911, reprinted 1974), still the best book on the subject; M. CARY, *A History of the Greek World from 323 to 146 B.C.*, 2nd ed., rev. (1951, reissued 1972), a useful and clear coverage of the chosen period; and M. ROSTOVTZEFF, *The Social and Economic History of the Hellenistic World*, 3 vol. (1941, reprinted 1986), a comprehensive and authoritative study, though at times controversial.

Particular topics are addressed by the following studies: H. IDRIS BELL, *Egypt, from Alexander the Great to the Arab Conquest: A Study in the Diffusion and Decay of Hellenism* (1948, reprinted 1977), a standard history by a great authority; EDWIN ROBERT BEVAN, *The House of Seleucus*, 2 vol. (1902, reprinted 2 vol. in 1, 1985), still the best treatment in English; ESTHER V. HANSEN, *The Attalids of Pergamon*, 2nd ed., rev. and expanded (1971), a major original synthesis; W.W. TARN, *The Greeks in Bactria & India*, 3rd ed. updated by FRANK LEE HOLT (1985), a pioneering and controversial work; A.K. NARAIN, *The Indo-Greeks* (1957, reissued 1980), critical and sympathetic; and DAVID MAGIE, *Roman Rule in Asia Minor: To the End of the Third Century after Christ*, 2 vol. (1950, reprinted 1988).

Hellenistic culture and science are described and discussed by T.B.L. WEBSTER, *Hellenistic Art* (1967), a general work; CHRISTINE MITCHELL HAVELOCK, *Hellenistic Art*, 2nd ed. (1981), detailed studies of individual items; MARGARETE BIEBER, *The Sculpture of the Hellenistic Age*, 2nd rev. ed. (1981), well-illustrated and reliable; J. CHARBONNEAUX, R. MARTIN, and F. VILLARD, *Hellenistic Art, 330–50 BC* (1973, originally published in French, 1970), magisterial and richly illustrated; ALBIN LESKY, *A History of Greek Literature* (1966; originally published in German, 2nd ed., 1963), brilliant, but only partly on the period; GEORGE SARTON, *A History of Science*, vol. 2, *Hellenistic Science and Culture in the Last Three Centuries B.C.* (1959), an authoritative summary; E.D. PHILLIPS, *Greek Medicine* (1973, reissued as *Aspects of Greek Medicine*, 1987), a good overview; A.A. LONG, *Hellenistic Philosophy: Stoics, Epicureans, Sceptics*, 2nd ed. (1986), an excellent introduction; and MARTIN HENGEL, *Judaism and Hellenism: Studies in their Encounter in Palestine During the Early Hellenistic Period* (1974, reissued 1981; originally published in German, 1969), a magisterial study. Further studies on Hellenistic philosophy and on Hellenistic religions may be found in the bibliographies of the articles GREEK LITERATURE and EUROPEAN RELIGIONS, ANCIENT. (J.Fe.)

Ancient Italic peoples. The basic guide is *Popoli e civiltà dell'Italia antica*, 7 vol. (1974–78). Outdated but still useful are JOSHUA WHATMOUGH, *The Foundations of Roman Italy* (1937, reprinted 1971); and DAVID RANDALL-MACIVER, *Italy Before the Romans* (1928, reprinted 1972). On the Etruscans there are three essential books in English that guide the reader through the bewildering maze of recent discoveries and research: MASSIMO PALLOTTINO, *The Etruscans*, rev. and enlarged ed. edited by DAVID RIDGWAY (1975; originally published in Italian, 6th ed., rev. and enlarged, 1975); DAVID RIDGWAY and FRANCESCA R. RIDGWAY (eds.), *Italy Before the Romans: The Iron Age, Orientalizing, and Etruscan Periods* (1979); and LARISSA BONFANTE (ed.), *Etruscan Life and Afterlife: A Handbook of Etruscan Studies* (1986), which lists the many catalogs of exhibitions and other publications that came out as a result of "The Year of the Etruscans" in Italy in 1985. Among recent interpretations of Etruscan culture and history, the best in English are MAURO CRISTOFANI, *The Etruscans: A New Investigation* (1979; originally published in Italian, 1978); and MICHAEL GRANT, *The Etruscans* (1980). GIULIANO BONFANTE and LARISSA BONFANTE, *The Etruscan Language* (1983), is a helpful introduction. Works on other individual Italic peoples, civilizations, and languages include the Ridgways' book, cited earlier; LUIGI BERNABÒ BREA, *Sicily Before the Greeks*, rev. ed. (1966); RENATO PERONI, *Archeologia della Puglia preistorica* (1967); A. ALFÖLDI, *Early Rome and the Latins* (1963); GABRIELLA GIACOMELLI, *La lingua falisca* (1963); GIACOMO DEVOTO, *Gli antichi Italici*, 3rd ed. rev. (1967); E.T. SALMON, *Samnium and the Samnites* (1967); G.A. MANSUELLI and R. SCARANI, *L'emilia prima dei Romani* (1961); G.B. PELLEGRINI and A.L. PROSDOCIMI, *La lingua venetica*, 2 vol. (1967); and on Italic inscriptions, ORONZO PARLANGÈLI, *Studi messapici* (1960); and ALLESSANDRO MORANDI, *Epigrafia italica* (1982). (N.T. de G.)

Ancient Rome. *Rome from its origins to 264:* Archaeological evidence on early Rome is discussed and analyzed by RAYMOND BLOCH, *The Origins of Rome*, rev. ed. (1963; originally published in French, 1946); T.J. CORNELL, "Rome and Latium Vetus," *Archaeological Reports*, 26:71–88 (1979–80); and ROBERT DREWS, "The Coming of the City to Central Italy," *American Journal of Ancient History*, 6:133–165 (1981). The archaeology of early Italy in general is covered in DAVID TRUMP, *Central and Southern Italy Before Rome* (1966). Livy's work on early Rome is carefully annotated and commented on in part by R.M. OGILVIE, *A Commentary on Livy, Books 1–5* (1965, reissued 1984). A good survey of Livy's annalistic predecessors is E. BADIAN, "The Early Historians," in T.A. DOREY (ed.), *Latin Historians* (1966), pp. 1–38. The single best modern treatment of the regal period and the early republic is JACQUES HEURGON, *The Rise of Rome to 264 B.C.* (1973; originally published in French, 1969). A complete chronological listing of all known magistrates of the Roman Republic with full ancient citations can be found in T. ROBERT S. BROUGHTON, *The Magistrates of the Roman Republic*, 2 vol. and a supplement (1951–60, reprinted 1984–86). A collection and modern analysis of ancient sources concerning Rome's economic development is TENNEY FRANK (ed.), *An Economic Survey of Ancient Rome*, 6 vol. (1933–40, reprinted 1975). The legal evidence from early Rome is treated by ALAN WATSON, *Rome of the XII Tables: Persons and Property* (1975).

The evolution of Rome's foundation myth is discussed by E.J. BICKERMAN, "Origines Gentium," *Classical Philology*, 47(2):65–81 (April 1952). Bickerman treats a number of important methodological questions on early Rome in "Some Reflections on Early Roman History," *Rivista di Filologia e di Istruzione Classica*, 97:393–408 (1969). RICHARD I. RIDLEY, "Fastenkritik: A Stocktaking," *Athenaeum*, 58(3–4):264–298 (1980), surveys various modern views on the reliability of the consular *fasti*. The single best treatment of the Roman ruling class is MATTHIAS GELZER, *The Roman Nobility* (1969; originally published in German, 1912). The Roman assemblies and voting procedures are thoroughly examined by GEORGE WILLIS BOTSFORD, *The Roman Assemblies from Their Origin to the End of the Republic* (1909, reprinted 1968); and LILY ROSS TAYLOR, *Roman Voting Assemblies from the Hannibalic War to the Dictatorship of Caesar* (1966). Taylor has also carefully studied the origin and development of the 35 urban and rural voting tribes in *The Voting Districts of the Roman Republic* (1960). E. STUART STAVELEY, "Forschungsbericht: The Constitution of the Roman Republic 1940–1954," *Historia*, 5:74–122 (1956), surveys modern scholarship on a number of important constitutional problems of early Roman history. Staveley has discussed the problem of the distinction between patricians and plebeians in "The Nature and Aims of the Patriciate," *Historia*, 32:24–57 (1983). A collection of essays by different scholars addressing this same problem is KURT A. RAAFLAUB (ed.), *Social Struggles in Archaic Rome: New Perspectives on the Conflict of the Orders* (1986), which contains an excellent bibliography on early Rome. A detailed and novel approach to the problem of patricians and plebeians is RICHARD E. MITCHELL, *Patricians and Plebeians: The Origin of the Roman State* (1990). The single best treatment of the military tribunes with consular power and related questions is KURT VON FRITZ, "The Reorganization of the Roman Government in 366 B.C. and the So-called Licinio-Sextian Laws," *Historia*, 1:3–44 (1950).

The best modern discussion of Roman imperialism is WILLIAM V. HARRIS, *War and Imperialism in Republican Rome, 327–70 B.C.* (1979). Harris' *Rome in Etruria and Umbria* (1971), examines Rome's relations with those two regions. Other informative works on Roman expansion include R.M. ERRINGTON, *The Dawn of Empire: Rome's Rise to World Power* (1971); ERICH S. GRUEN, *The Hellenistic World and the Coming of Rome*, 2 vol. (1984); and E. BADIAN, *Foreign Clientelae, 264–70 B.C.* (1958). E.T. SALMON, *Roman Colonization Under the Republic* (1969), surveys the methods, aims, and consequences of Roman colonization. (G.E.Fo.)

The middle republic (264–133 BC): H.H. SCULLARD, *A History of the Roman World: 753–146 BC*, 4th ed. (1980), provides a reliable narrative. GAETANO DE SANCTIS, *Storia dei Romani*, 4 vol. (1907–65), is more detailed. The standard reference work on Polybius is F.W. WALBANK, *A Historical Commentary on Polybius*, 3 vol. (1957–79). On the wars with Carthage, ULRICH SCHREDT, *Geschichte der Karthager von 218–146* (1913, reprinted 1975), provides source criticism. Military aspects of this period are presented in JOHANNES KROMAYER and GEORG VEITH, *Antike Schlachtfelder*, vol. 3 in 2 parts (1912); JOHANNES KROMAYER and GEORG VEITH (eds.), *Schlachten-Atlas zur antiken Kriegsgeschichte*, 5 parts (1922–29); J.H. THIEL, *A History of Roman Sea-power Before the Second Punic War* (1954), and *Studies on the History of Roman Sea-power in Republican Times* (1946); J.F. LAZENBY, *Hannibal's War: A Military History of the Second Punic War* (1978); and H.H. SCULLARD, *Scip-*

to Africanus: Soldier and Politician (1970). STÉPHANE GSELL, *Histoire ancienne de l'Afrique du Nord*, 3rd ed., 8 vol. (1928); and B.H. WARMINGTON, *Carthage*, rev. ed. (1969), deal with Carthage. Works on the provinces include DAVID MAGIE, *Roman Rule in Asia Minor to the End of the Third Century After Christ*, 2 vol. (1950, reissued 1988); G.H. STEVENSON, *Roman Provincial Administration till the Age of the Antonines* (1939, reprinted 1975); and C.H.V. SUTHERLAND, *The Romans in Spain, 217 B.C.–A.D. 117* (1939, reprinted 1982).

The transformation of Rome and Italy during the middle republic: Citizenship, constitution, and politics are discussed in THEODOR MOMMSEN, *Römisches Staatsrecht*, 3rd ed., 3 vol. in 5 (1887–88, reprinted 1969); A.N. SHERWIN-WHITE, *The Roman Citizenship*, 2nd ed. (1973, reissued 1987); and C. NICOLET, *The World of the Citizen in Republican Rome* (1980; originally published in French, 1976). ARNOLD J. TOYNBEE, *Hannibal's Legacy: The Hannibalic War's Effects on Roman Life*, 2 vols. (1965); P.A. BRUNT, *Italian Manpower, 225 B.C.–A.D. 14* (1971, reprinted 1987); and KEITH HOPKINS, *Conquerors and Slaves* (1977), explore the social and economic consequences of Rome's victories. P.A. BRUNT, *Social Conflicts in the Roman Republic* (1971, reissued 1986), presents an excellent brief account. Many important aspects of second-century politics and culture are covered in ALAN E. ASTIN, *Scipio Aemilianus* (1967).

The late republic (133–31 BC): The best outline in English for the late republic is the first half of H.H. SCULLARD, *From the Gracchi to Nero*, 5th ed. (1982), with excellent notes and bibliography. The classic reference work is W. DRUMANN, *Geschichte Roms in seinem Übergange von der republikanischen zur monarchischen Verfassung*, 2nd ed. edited by P. GROEBE, 6 vol. (1899–1929), giving biographies (with full source material) of all prominent figures of the period, arranged by families. Classic interpretations of the fall of the republic are RONALD SYME, *The Roman Revolution* (1939, reissued 1987); P.A. BRUNT, *The Fall of the Roman Republic and Related Essays* (1988); LILY ROSS TAYLOR, *Party Politics in the Age of Caesar* (1949, reissued 1975); ERICH S. GRUEN, *The Last Generation of the Roman Republic* (1974); and MATTHIAS GELZER, *Caesar: Politician and Statesman* (1968; originally published in German, 1940). The army and expansion are analyzed in EMILIO GABBA, *Republican Rome, the Army, and the Allies* (1976; originally published in Italian, 1973); and E. BADIAN, *Roman Imperialism in the Late Republic*, 2nd ed. (1968). Aspects of public and social life are dealt with in T.P. WISEMAN, *New Men in the Roman Senate, 139 B.C.–A.D. 14* (1971); ISRAËL SHATZMAN, *Senatorial Wealth and Roman Politics* (1975); SUSAN TREGGIARI, *Roman Freedmen During the Late Republic* (1969); A.W. LINTOTT, *Violence in Republican Rome* (1968); and E. BADIAN, *Publicans and Sinners: Private Enterprise in the Service of the Roman Republic* (1972, reissued 1983). On cultural development, the standard work is ELIZABETH RAWSON, *Intellectual Life in the Late Roman Republic* (1985); it may be supplemented by J.H.W.G. LIEBESCHUETZ, *Continuity and Change in Roman Religion* (1979); BRUCE W. FRIER, *The Rise of the Roman Jurists* (1985); and GEORGE KENNEDY, *The Art of Rhetoric in the Roman World, 300 B.C.–A.D. 300* (1972). (R.P.Sa.)

The early Roman Empire (31 BC–AD 193): COLIN WELLS, *The Roman Empire* (1984), is an intelligent short history up through the Severi. The history is carried further by MICHAEL GRANT, *The Climax of Rome: The Final Achievements of the Ancient World, A.D. 161–337* (1968). DONALD EARL, *The Age of Augustus* (1968, reissued 1980), is useful in providing a little more depth. As to governmental institutions, FERGUS MILLAR, *The Emperor in the Roman World, 31 BC–AD 337* (1977), offers a monumentally detailed study of the ruler in his capacity as civil governor up through Constantine; and RICHARD J.A. TALBERT, *The Senate of Imperial Rome* (1984), describes the role and actions of the ruler's partner. On provincial government, as well as much else, FERGUS MILLAR (ed.), *The Roman Empire and Its Neighbours*, 2nd ed. (1981; originally published in German, 1966), is informative and readable. Commentary on the economy is supplied by KEVIN GREENE, *The Archaeology of the Roman Economy* (1986). GÉZA ALFÖLDY, *The Social History of Rome* (1985), on the structure of society; and RAMSAY MACMULLEN, *Roman Social Relations, 50 B.C. to A.D. 284* (1974), on the feelings uniting or dividing groups or strata, are complementary works. Provincial history broadly interpreted may be sampled in SHEPPARD FREERE, *Britannia: A History of Roman Britain*, 3rd ed. rev. (1987); PAUL MACKENDRICK, *The North African Stones Speak* (1980); and A.H.M. JONES, *The Greek City from Alexander to Justinian* (1940, reissued 1979), still useful, since archaeology has little touched the eastern end of the Mediterranean world. BERNARD ANDREAE, *The Art of Rome* (1977; originally published in German, 1973), a large, luxuriously illustrated work with an equally rich scholarly text; and NIELS HANNESTAD, *Roman Art and Imperial Policy* (1986; originally published in Danish, 1976), deal with their material in quite different ways: the former is conventionally art-historical, the latter uses his material to illuminate its context. Architecture is best approached through W.L. MACDONALD, *The Architecture of the Roman Empire*, rev. ed., 2 vol. (1982–86), a well-written, imaginative account; and through such specialized studies as JOHN PERCIVAL, *The Roman Villa: An Historical Introduction* (1976, reissued 1988). PHILIPPE ARIÈS and GEORGES DUBY (eds.), *A History of Private Life*, vol. 1, *From Pagan Rome to Byzantium*, ed. by PAUL VEYNE (1987; originally published in French, 1985), is a social history in an old-fashioned sense by a master of the most up-to-date approaches. The importance of emperor worship is well argued in the detailed work by DUNCAN FISHWICK, *The Imperial Cult in the Latin West*, vol. 1 in 2 vol. (1987); and, with more interpretation and for the other half of the empire, by S.R.F. PRICE, *Rituals and Power: The Roman Imperial Cult in Asia Minor* (1984). RAMSAY MACMULLEN, *Paganism in the Roman Empire* (1981), provides a comprehensive view. Military history is made accessible through G.R. WATSON, *The Roman Soldier* (1969, reissued 1985). An explication of a major aspect of culture may be found in the latter half of a work by a notable historian, H.I. MARROU, *A History of Education in Antiquity* (1956, reprinted 1982; originally published in French, 1948). ALBIN LESKY, *A History of Greek Literature* (1966; originally published in German, 2nd ed., 1963), may be paired with H.J. ROSE, *A Handbook of Latin Literature, from the Earliest Times to the Death of St. Augustine*, 3rd ed. (1966); and with the more elegant study by GORDON WILLIAMS, *Change and Decline: Roman Literature in the Early Empire* (1978). On the church, W.H.C. FREND, *The Rise of Christianity* (1984), is readable and comprehensive up through the 6th century.

The later Roman Empire: ANDRÉ PIGANIOL, *L'Empire chrétien (325–395)*, 2nd ed. updated by ANDRÉ CHASTAGNOL (1972), offers an exceptionally rich and informative narrative among modern works. DIANA BOWDER, *The Age of Constantine and Julian* (1978), is good on those two reigns. A.H.M. JONES, *The Later Roman Empire, 284–602: A Social Economic and Administrative Survey*, 2 vol. (1964, reprinted 1986), is extraordinarily clear and detailed on these topics. On a major development, monasticism, DERWAS J. CHITTY, *The Desert a City: An Introduction to the Study of Egyptian and Palestinian Monasticism Under the Christian Empire* (1966, reissued 1977), is highly readable. RAMSAY MACMULLEN, *Corruption and the Decline of Rome* (1988), includes an up-to-date survey of evidence for decline, and also argues a thesis. HERWIG WOLFRAM, *History of the Goths* (1988; originally published in German, 2nd ed., 1980), is a superb study of a crucial player in the 4th to 6th centuries. WALTER GOFFART, *Barbarians and Romans, A.D. 418–584: The Techniques of Accommodation* (1980), carries the account further. (R.MacM.)

The Classical Greek Dramatists: Aeschylus, Sophocles, Euripides, and Aristophanes

Of the literature of ancient Greece only a small portion survives. Yet much of it remains important, not only because of its high quality but also because Western literature is partly based on forms that were of Greek invention. Two such literary forms are tragedy (*tragōidia;* "goat-song") and comedy (*komōidia;* "revel-song"), both of which presumably originated in ancient Greek rituals of celebration and sacrifice. Both tragedy and comedy were performed at Athens at the spring festival of the god Dionysus (the Great Dionysia). The only extant Greek tragedies are a total of 35 plays written by the Athenian poet-playwrights Aeschylus, Sophocles, and Euripides. In the works of these three men Greek tragedy can be seen both at its creation and at the height of its formal and artistic development. Similarly, the 11 extant complete plays by the Athenian playwright Aristophanes are the only surviving examples of the very first stage in the development of comedy in the Western world, known as Old Comedy. This article treats the lives, works, and achievements of these four classical dramatists. For treatment of the origins, forms, and development of ancient Greek drama and dramatic production, see GREEK LITERATURE; THEATRE, THE HISTORY OF WESTERN; THEATRICAL PRODUCTION.

The article is divided into the following sections:

Aeschylus 342
Life and career 342
Dramatic and literary achievements 342
The plays 343
Sophocles 344
Life and career 344
Dramatic and literary achievements 344
The plays 345
Euripides 346
Life and career 346
Dramatic and literary achievements 346
The plays 347
Aristophanes 348
Life and career 349
Dramatic and literary achievements 349
The plays 349
Major works 350
Bibliography 350

Aeschylus

LIFE AND CAREER

Aeschylus was the first of classical Athens' three great writers of tragedy. He grew up in the turbulent period when the Athenian democracy, having thrown off its tyranny (the absolute rule of one man), had to prove itself against both self-seeking politicians at home and invaders from abroad. Aeschylus himself took part in his city's first struggles against the invading Persians. Later Greek chroniclers believed that Aeschylus was 35 years old in 490 BC when he participated in the Battle of Marathon, in which the Athenians first repelled the Persians; if this is true it would place his birth in 525 BC. Aeschylus' father's name was Euphorion, and the family probably lived at Eleusis (west of Athens).

Aeschylus was a notable participant in Athens' major dramatic competition, the Great Dionysia, which was a part of the festival of Dionysus. Every year at this festival, each of three dramatists would produce three tragedies, which either could be unconnected in plot sequence or could have a connecting theme. This trilogy was followed by a satyr play, which was a kind of lighthearted burlesque. Aeschylus is recorded as having participated in this competition, probably for the first time, in 499 BC. He won his first victory in the theatre in the spring of 484 BC. In the meantime, he had fought and possibly been wounded at Marathon, and Aeschylus singled out his participation in this battle years later for mention on the verse epitaph he wrote for himself. Aeschylus' brother was killed in this battle. In 480 the Persians again invaded Greece, and once again Aeschylus saw service, fighting at the battles of Artemisium and Salamis. His responses to the Persian invasion found expression in his play *Persians,* the earliest of his works to survive. This play was produced in the competition of the spring of 472 BC and won first prize.

Around this time Aeschylus is said to have visited Sicily to present *Persians* again at the tyrant Hieron I's court in Syracuse. Aeschylus' later career is a record of sustained dramatic success, though he is said to have suffered one memorable defeat, at the hands of the novice Sophocles, whose entry at the Dionysian festival of 468 BC was victorious over the older poet's entry. Aeschylus recouped the loss with victory in the next year, 467, with his Oedipus trilogy (of which the third play, *Seven Against Thebes,* survives). After producing his masterpiece, the *Oresteia* trilogy, in 458, Aeschylus went to Sicily again. The chronographers recorded Aeschylus' death at Gela (on Sicily's south coast) in 456/455, aged 69. A ludicrous story that he was killed when an eagle dropped a tortoise on his bald pate was presumably fabricated by a later comic writer. At Gela he was accorded a public funeral, with sacrifices and dramatic performances held at his grave, which subsequently became a place of pilgrimage for writers.

Aeschylus wrote approximately 90 plays, including satyr plays as well as tragedies; of these, about 80 titles are known. Only seven tragedies have survived entire. One account, perhaps based on the official lists, assigns Aeschylus 13 first prizes, or victories; this would mean that well over half of his plays won, since sets of four plays rather than separate ones were judged. According to the philosopher Flavius Philostratus, Aeschylus was known as the "Father of Tragedy." Aeschylus' two sons also achieved prominence as tragedians. One of them, Euphorion, won first prize in his own right in 431 BC over Sophocles and Euripides.

DRAMATIC AND LITERARY ACHIEVEMENTS

Aeschylus' influence on the development of tragedy was fundamental. Previous to him, Greek drama was limited to one actor and a chorus engaged in a largely static recitation. (The chorus was a group of actors who responded to and commented on the main action of a play with song, dance, and recitation.) The actor could assume different roles by changing masks and costumes, but he was limited to engaging in dialogue only with the chorus. By adding a second actor with whom the first could converse, Aeschylus vastly increased the drama's possibilities for dialogue and dramatic tension and allowed more variety and freedom in plot construction. Although the dominance of the chorus in early tragedy is ultimately only hypothesis, it is probably true that, as Aristotle says in his *Poetics,* Aeschylus "reduced the chorus' role and made the plot the leading actor." Aeschylus was an innovator in other ways as well. He made good use of stage settings and stage machinery, and some of his works were noted for their spectacular scenic effects. He also designed costumes, trained his choruses in their songs and dances, and probably acted in most of his own plays, this being the usual practice among Greek dramatists.

Formal innovations

But Aeschylus' formal innovations account for only part of his achievement. His plays are of lasting literary value

in their majestic and compelling lyrical language, in the intricate architecture of their plots, and in the universal themes which they explore so honestly. Aeschylus' language in both dialogue and choral lyric is marked by force, majesty, and emotional intensity. He makes bold use of compound epithets, metaphors, and figurative turns of speech, but this rich language is firmly harnessed to the dramatic action rather than used as mere decoration. It is characteristic of Aeschylus to sustain an image or group of images throughout a play; the ship of state in *Seven Against Thebes,* the birds of prey in *Suppliants,* the snare in *Agamemnon.* More generally, Aeschylus deploys throughout a play or trilogy of plays several leading motifs that are often associated with a particular word or group of words. In the *Oresteia,* for example, such themes as wrath, mastery, persuasion, and the contrasts of light and darkness, of dirge and triumphal song, run throughout the trilogy. This sort of dramatic orchestration as applied to careful plot construction enabled Aeschylus to give Greek drama a more truly artistic and intellectual form.

Aeschylean tragedy deals with the plights, decisions, and fates of individuals with whom the destiny of the community or state is closely bound up; in turn, both individual and community stand in close relation to the gods. Personal, social, and religious issues are thus integrated, as they still were in the Greek civilization of the poet's time. Theodicy (*i.e.,* the justifying of the gods' ways to men) was in some sense the concern of Aeschylus, though it might be truer to say that he aimed through dramatic conflict to throw light on the nature of divine justice. Aeschylus and his Greek contemporaries believed that the gods begrudged human greatness and sent infatuation on a man at the height of his success, thus bringing him to disaster. Man's infatuated act was frequently one of impiety or pride (*hubris*), for which his downfall could be seen as a just punishment. In this scheme of things, divine jealousy and eternal justice formed the common fabric of a moral order of which Zeus, supreme among the gods, was the guardian.

But the unjust are not always punished in their lifetime; it is upon their descendants that justice may fall. It was this tradition of belief in a just Zeus and in hereditary guilt that Aeschylus received, and which is evinced in many of his plays. The simplest illustration of this is in *Persians,* in which Xerxes and his invading Persians are punished for their own offenses. But in a play such as *Agamemnon,* the issues of just punishment and moral responsibility, of human innocence and guilt, of individual freedom versus evil heredity and divine compulsion are more complex and less easily disentangled, thus presenting contradictions which still baffle the human intellect.

Theme of divine justice

Finally, to Aeschylus, divine justice uses human motives to carry out its decrees. Chief among these motives is the desire for vengeance, which was basic to the ancient Greek scheme of values. In the one complete extant trilogy, the *Oresteia,* this notion of vengeance or retaliation is dominant. Retaliation is a motive of Agamemnon, Clytemnestra, Aegisthus, and Orestes. But significantly, the chain of retaliatory murder that pursues Agamemnon and his family ends not by a perfect balance of blood guilt, not by a further perpetuation of violence, but rather through reconciliation and the rule of law as established by Athena and the Athenian courts of justice.

Aeschylus is almost unequaled in writing tragedy that, for all its power of depicting evil and the fear and consequences of evil, ends, as in the *Oresteia,* in joy and reconciliation. Living at a time when the Greek people still truly felt themselves surrounded by the gods, Aeschylus nevertheless had a capacity for detached and general thought, which was typically Greek and which enabled him to treat the fundamental problem of evil with singular honesty and success.

THE PLAYS

Persians. One of a trilogy of unconnected tragedies presented in 472 BC, *Persians* (Greek *Persai*) is unique among surviving tragedies in that it dramatizes recent history rather than events from the distant age of mythical heroes. The play treats the decisive repulse of the Persians from Greece in 480, in particular their defeat at the Battle of Salamis. The play is set in the Persian capital, where a messenger brings news to the Persian queen of the disaster at Salamis. After attributing the defeat of Persia to both Greek independence and bravery and to the gods' punishment of Persian folly for going outside the bounds of Asia, the play ends with the return of the broken and humiliated Persian king, Xerxes.

Seven Against Thebes. This is the third and only surviving play of a connected trilogy, presented in 467 BC, that dealt with the impious transgressions of Laius and the doom subsequently inflicted upon his descendants. The first play seems to have shown how Laius, king of Thebes, had a son despite the prohibition of the oracle of the god Apollo. In the second play it appears that that son, Oedipus, killed his father and laid a curse on his own two sons, Eteocles and Polyneices. In *Seven Against Thebes* (Greek *Hepta epi Thēbais*) Eteocles is shown leading the defense of the city of Thebes against an invading army led by his brother Polyneices and six chieftains from the south of Greece who are bent on placing Polyneices on the Theban throne. Eteocles assigns defenders to each of six of the seven gates of Thebes; but he insists on fighting at the seventh gate, where his opponent will be Polyneices. There the brothers kill each other, and the Theban royal family is thus exterminated, bringing to an end the horrors set in motion by Laius' defiance of the gods.

Suppliants. This is the first and only surviving play of a trilogy probably put on in 463. It was long believed by scholars that *Suppliants* (Greek *Hiketides;* Latin *Supplices*) was one of Aeschylus' earliest plays because of its archaic structure; its chorus, representing the daughters of Danaus (the Danaids), takes the leading role in the action. But there is now evidence that the trilogy of which *Suppliants* formed a part was produced in competition with Sophocles, who is first known to have competed in 468. *Suppliants* thus dates presumably from the middle of Aeschylus' career, not from the beginning.

Born in Egypt, though of Greek descent, the Danaïds have fled with their father to Argos in Greece in order to avoid forced marriage with their cousins, the sons of Aegyptus. Pelasgus, the king of Argos, is torn between charity to the Danaïds and anxiety to appease Aegyptus but nobly agrees in the end to grant them asylum. The trilogy as a whole seems to have favourably stressed the saving power of domestic love as contrasted with both the willful virginity of the Danaïds and the unfeeling, violent lust of their cousins.

Oresteia. The *Oresteia* trilogy consists of three closely connected plays, all extant, that were presented in 458 BC. In *Agamemnon* the great Greek king of that name returns triumphant from the siege of Troy, along with his concubine, the Trojan prophetess Cassandra, only to be humiliated and murdered by his fiercely vengeful wife, Clytemnestra. She is driven to this act partly by a desire to avenge the death of her daughter Iphigenia, whom Agamemnon has sacrificed for the sake of the war, partly by her adulterous love for Aegisthus, and partly as agent for the curse brought on Agamemnon's family by the crimes of his father, Atreus. At the play's end Clytemnestra and her lover have taken over the palace and now rule Argos. Many regard this play as one of the greatest Greek tragedies. From its extraordinarily sustained dramatic and poetic power one might single out the fascinating, deceitful richness of Clytemnestra's words and the huge choral songs, which raise in metaphorical and often enigmatic terms the complex of major themes—of theology, politics, and blood relationships—which are elaborated throughout the trilogy.

Oresteia trilogy

Libation Bearers (Greek *Choēphoroi*) is the second play in the trilogy and takes its title from the chorus of women servants who come to pour propitiatory offerings at the tomb of the murdered Agamemnon. At the start of this play Orestes, the son of Agamemnon and Clytemnestra, who was sent abroad as a child, returns as a man to take vengeance upon his mother and her lover for their murder of his father. He is reunited with his sister Electra, and together they invoke the aid of the dead Agamemnon in their plans. Orestes then slays Aegisthus, but Orestes'

subsequent murder of Clytemnestra is committed reluctantly, at the god Apollo's bidding. Orestes' attempts at self-justification then falter and he flees, guilt-wracked, maddened, and pursued by the female incarnations of his mother's curse, the Erinyes (Furies). At this point the chain of vengeance seems interminable.

Eumenides, the title of the third play, means "The Kind Goddesses." The play opens at the shrine of Apollo at Delphi, where Orestes has taken sanctuary from the Furies. At the command of the Delphic oracle, Orestes journeys to Athens to stand trial for his matricide. There the goddess Athena organizes a trial with a jury of citizens. The Furies are his accusers, while Apollo defends him. The jury divides evenly in its vote and Athena casts the tie-breaking vote for Orestes' acquittal. The Furies then turn their vengeful resentment against the city itself, but Athena persuades them, in return for a home and cult, to bless Athens instead and reside there as the "Kind Goddesses" of the play's title. The trilogy thus ends with the cycle of retributive bloodshed ended and supplanted by the rule of law and the justice of the state.

Prometheus Bound. The date of this play (and even its authorship) is disputed, but many scholars regard it as a work of Aeschylus' last years. In *Prometheus Bound* (Greek *Promētheus desmōtēs*) the god Prometheus, who in defiance of Zeus has saved mankind and given them fire, is chained to a remote crag as a punishment ordered by the king of the gods. Despite his isolation Prometheus is visited by the ancient god Oceanus, by a chorus of Oceanus' daughters, by the "cow-headed" Io (another victim of Zeus), and finally by the god Hermes, who vainly demands from Prometheus his knowledge of a secret that could threaten Zeus's power. After refusing to reveal his secret, Prometheus is cast into the underworld for further torture. The drama of the play lies in the clash between the irresistible power of Zeus and the immovable will of Prometheus, who has been rendered still more stubborn by Io's misfortunes at the hands of Zeus. The most striking and controversial aspect of the play is its depiction of Zeus as a tyrant. Prometheus himself has proved to be for later ages an archetypal figure of defiance against tyrannical power, a role exemplified in Percy Bysshe Shelley's poem *Prometheus Unbound* (1820).

(A.J.P./O.T.)

Sophocles

LIFE AND CAREER

The second of classical Athens' three great writers of tragedy, Sophocles was the younger contemporary of Aeschylus and the older contemporary of Euripides. He was born about 496 BC at Colonus, a village outside the walls of Athens. His father, Sophillus, was a wealthy manufacturer of armour, and Sophocles himself received a good education. Because of his beauty of physique, his athletic prowess, and his skill in music, he was chosen in 480, when he was 16, to lead the paean (choral chant to a god) celebrating the decisive Greek sea victory over the Persians at the Battle of Salamis. The relatively meagre information about Sophocles' civic life suggests that he was a popular favourite who participated actively in his community and exercised outstanding artistic talents. In 442 he served as one of the treasurers responsible for receiving and managing tribute money from Athens' subject-allies in the Delian League. In 440 he was elected one of the 10 *stratēgoi* (high executive officials who commanded the armed forces) as a junior colleague of Pericles. Sophocles later served as *stratēgos* perhaps twice again. In 413, then aged about 83, Sophocles was a *proboulos,* one of 10 advisory commissioners who were granted special powers and were entrusted with organizing Athens' financial and domestic recovery after its terrible defeat at Syracuse in Sicily. Sophocles' last recorded act was to lead a chorus in public mourning for his deceased rival, Euripides, before the festival of 406. He died that same year.

Public service

These few facts are about all that is known of Sophocles' life. They imply steady and distinguished attachment to Athens, its government, religion, and social forms. Sophocles was wealthy from birth, highly educated, noted for his grace and charm, on easy terms with the leading families, a personal friend of prominent statesmen, and in many ways fortunate to have died before the final surrender of Athens to Sparta in 404. In one of his last plays, *Oedipus at Colonus,* he still affectionately praises both his own birthplace and the great city itself.



Sophocles, bronze bust copied from a Greek original, 340–330 BC. In the Museo Archeologico, Florence.
Alinari—Art Resource/EB Inc.

Sophocles won his first victory at the Dionysian dramatic festival in 468, however, defeating the great Aeschylus in the process. This began a career of unparalleled success and longevity. In total, Sophocles wrote 123 dramas for the festivals. Since each author who was chosen to enter the competition usually presented four plays, this means he must have competed about 30 times. Sophocles won perhaps as many as 24 victories, compared to 13 for Aeschylus and four for Euripides, and indeed he may have never received lower than second place in the competitions he entered.

DRAMATIC AND LITERARY ACHIEVEMENTS

Ancient authorities credit Sophocles with several major and minor dramatic innovations. Among the latter is his invention of some type of "scene paintings" or other pictorial prop to establish locale or atmosphere. He also may have increased the size of the chorus from 12 to 15 members. Sophocles' major innovation was his introduction of a third actor into the dramatic performance. It had previously been permissible for two actors to "double" (*i.e.,* assume other roles during a play), but the addition of a third actor onstage enabled the dramatist both to increase the number of his characters and widen the variety of their interactions. The scope of the dramatic conflict was thereby extended, plots could be more fluid, and situations could be more complex.

The typical Sophoclean drama presents a few characters, impressive in their determination and power and possessing a few strongly drawn qualities or faults that combine with a particular set of circumstances to lead them inevitably to a tragic fate. Sophocles develops his characters' rush to tragedy with great economy, concentration, and dramatic effectiveness, creating a coherent, suspenseful situation whose sustained and inexorable onrush came to epitomize the tragic form to the classical world. Sophocles emphasizes that most people lack wisdom, and he presents truth in collision with ignorance, delusion, and folly. Many scenes dramatize flaws or failure in thinking (deceptive reports and rumours, false optimism, hasty judgment,

Typical dramatic framework

madness). The chief character does something involving grave error; this affects others, each of whom reacts in his own way, thereby causing the chief agent to take another step toward ruin—his own and that of others as well. Equally important, those who are to suffer from the tragic error usually are present at the time or belong to the same generation. It was this more complex type of tragedy that demanded a third actor. Sophocles thus abandoned the spacious Aeschylean framework of the connected trilogy and instead comprised the entire action in a single play. From his time onward, "trilogy" usually meant no more than three separate tragedies written by the same author and presented at the same festival.

Sophocles' language responds flexibly to the dramatic needs of the moment; it can be ponderously weighty or swift-moving, emotionally intense or easygoing, highly decorative or perfectly plain and simple. His mastery of form and diction was highly respected by his contemporaries. Sophocles has also been universally admired for the sympathy and vividness with which he delineates his characters; especially notable are his tragic women, such as Electra and Antigone. Few dramatists have been able to handle situation and plot with more power and certainty; the frequent references in the *Poetics* to Sophocles' *Oedipus the King* show that Aristotle regarded this play as a masterpiece of construction, and few later critics have dissented. Sophocles is also unsurpassed in his moments of high dramatic tension and in his revealing use of tragic irony.

The criticism has been made that Sophocles was a superb artist and nothing more; he grappled neither with religious problems as Aeschylus had nor with intellectual ones as Euripides had done. He accepted the gods of Greek religion in a spirit of unreflecting orthodoxy, and he contented himself with presenting human characters and human conflicts. But it should be stressed that to Sophocles "the gods" appear to have represented the natural forces of the universe to which human beings are unwittingly or unwillingly subject. To Sophocles, human beings live for the most part in dark ignorance because they are cut off from these permanent, unchanging forces and structures of reality. Yet it is pain, suffering, and the endurance of tragic crisis that can bring people into valid contact with the universal order of things. In the process, a person can become more genuinely human, more genuinely himself.

Conception of life

THE PLAYS

Only seven of Sophocles' tragedies survive in their entirety, along with 400 lines of a satyr play, numerous fragments of plays now lost, and 90 titles. All seven of the complete plays are works of Sophocles' maturity, but only two of them, *Philoctetes* and *Oedipus at Colonus*, have fairly certain dates. *Ajax* is generally regarded as the earliest of the extant plays. Some evidence suggests that *Antigone* was first performed in 442 or 441 BC. *Philoctetes* was first performed in 409, when Sophocles was 90 years old, and *Oedipus at Colonus* was said to have been produced after Sophocles' death by his grandson.

Ajax. The entire plot of *Ajax* (Greek *Aias mastigophoros*) is constructed around Ajax, the mighty hero of the Trojan War whose pride drives him to treachery and finally to his own ruin and suicide some two-thirds of the way through the play. Ajax is deeply offended at the award of the prize of valour (the dead Achilles' armour) not to himself but to Odysseus. Ajax thereupon attempts to assassinate Odysseus and the contest's judges, the Greek commanders Agamemnon and Menelaus, but is frustrated by the intervention of the goddess Athena. He cannot bear his humiliation and throws himself on his own sword. Agamemnon and Menelaus order that Ajax' corpse be left unburied as punishment. But the wise Odysseus persuades the commanders to relent and grant Ajax an honourable burial. In the end Odysseus is the only person who seems truly aware of the changeability of human fortune.

Antigone. Antigone is the daughter of Oedipus, the former king of Thebes. She is willing to face the capital punishment that has been decreed by her uncle Creon, the new king, as the penalty for anyone burying her brother Polyneices. (Polyneices has just been killed attacking Thebes, and it is as posthumous punishment for this attack that Creon has forbidden the burial of his corpse.) Obeying all her instincts of love, loyalty, and humanity, Antigone defies Creon and dutifully buries her brother's corpse. Creon, from conviction that reasons of state outweigh family ties, refuses to commute Antigone's death sentence. By the time Creon is finally persuaded by the prophet Tiresias to relent and free Antigone, she has killed herself in her prison cell. Creon's son, Haemon, kills himself out of love and sympathy for the dead Antigone, and Creon's wife, Eurydice, then kills herself out of grief over these tragic events. At the play's end Creon is left desolate and broken in spirit. In his narrow and unduly rigid adherence to his civic duties, Creon has defied the gods through his denial of humanity's common obligations toward the dead. The play thus concerns the conflicting obligations of civic versus personal loyalties and religious mores.

Trachinian Women. This play centres on the efforts of Deianeira to win back the wandering affections of her husband, Heracles, who is away on one of his heroic missions and who has sent back his latest concubine, Iole, to live with his wife at their home in Trachis. The love charm Deianeira uses on Heracles turns out to be poisonous, and she kills herself upon learning of the agony she has caused her husband. Thus, in *Trachinian Women* (Greek *Trachiniai*) Heracles' insensitivity (in sending his mistress to share his wife's home) and Deianeira's ignorance result in domestic tragedy.

Oedipus the King. The plot of *Oedipus the King* (Greek *Oidipous Tyrannos*; Latin *Oedipus Rex*) is a structural marvel that marks the summit of classical Greek drama's formal achievements. The play's main character, Oedipus, is the wise, happy, and beloved ruler of Thebes. Though hot-tempered, impatient, and arrogant at times of crisis, he otherwise seems to enjoy every good fortune. But Oedipus mistakenly believes that he is the son of King Polybus of Corinth and his queen. He became the ruler of Thebes because he rescued the city from the Sphinx by answering its riddle correctly, and so was awarded the city's widowed queen, Jocasta. Before overcoming the Sphinx, Oedipus left Corinth forever because the Delphic oracle had prophesied to him that he would kill his father and marry his mother. While journeying to Thebes from Corinth, Oedipus encountered at a crossroads an old man accompanied by five servants. Oedipus got into an argument with him and in a fit of arrogance and bad temper killed the old man and four of his servants.

The play opens with the city of Thebes stricken by a plague and its citizens begging Oedipus to find a remedy. He consults the Delphic oracle, which declares that the plague will cease only when the murderer of Jocasta's first husband, King Laius, has been found and punished for his deed. Oedipus resolves to find Laius' killer, and much of the rest of the play centres upon the investigation he conducts in this regard. In a series of tense, gripping, and ominous scenes Oedipus' investigation turns into an obsessive reconstruction of his own hidden past as he begins to suspect that the old man he killed at the crossroads was none other than Laius. Finally, Oedipus learns that he himself was abandoned to die as a baby by Laius and Jocasta because they feared a prophecy that their infant son would kill his father; that he survived and was adopted by the ruler of Corinth, but in his maturity he has unwittingly fulfilled the Delphic oracle's prophecy of him; that he has indeed killed his true father, married his own mother, and begot children who are also his own siblings.

Plot of *Oedipus the King*

Jocasta hangs herself when she sees this shameful web of incest, parricide, and attempted child murder, and the guilt-stricken Oedipus then sticks needles into his eyes, blinding himself. Sightless and alone, he is now blind to the world around him but finally cognizant of the terrible truth of his own life.

Electra. As in Aeschylus' *Libation Bearers*, the action in *Electra* (Greek *Elektra*) follows the return of Orestes to kill his mother, Clytemnestra, and her lover Aegisthus in retribution for their murder of Orestes' father, Agamemnon. In this play, however, the main focus is on Orestes' sister Electra and her anguished participation in her brother's plans. To gain admittance to the palace and thus be

Portrait of Electra

able to execute his revenge, Orestes spreads false news of his own death. Believing this report, the despairing Electra unsuccessfully tries to enlist her sister Chrysothemis in an attempt to murder their mother. In a dramatic scene, Orestes then enters in disguise and hands Electra the urn that is supposed to contain his own ashes. Moved by his sister's display of grief, Orestes reveals his true identity to her and then strikes down his mother and her lover. Electra's triumph is thus complete. In the play Electra is seen passing through the whole range of human emotions—from passionate love to cruel hatred, from numb despair to wild joy. There is debate over whether the play depicts virtue triumphant or, rather, portrays a young woman incurably twisted by years of hatred and resentment.

Philoctetes. In *Philoctetes* (Greek *Philoktētēs*) the Greeks on their way to Troy have cast away the play's main character, Philoctetes, on the desert island of Lemnos because he has a loathsome and incurable ulcer on his foot. But the Greeks have discovered that they cannot win victory over Troy without Philoctetes and his wonderful bow, which formerly belonged to Heracles. The crafty Odysseus is given the task of fetching Philoctetes by any means possible. Odysseus knows that the resentful Philoctetes will kill him if he can, so he uses the young and impressionable soldier Neoptolemus, son of the dead Achilles, as his agent. Neoptolemus is thus caught between the devious manipulations of Odysseus and the unsuspecting integrity of Philoctetes, who is ready to do anything rather than help the Greeks who abandoned him. For much of the play Neoptolemus sticks to Odysseus' policy of deceit, despite his better nature, but eventually he renounces duplicity to join in friendship with Philoctetes. A supernatural appearance by Heracles then convinces Philoctetes to go to Troy to both win victory and be healed of his disease.

Oedipus at Colonus. In *Oedipus at Colonus* (Greek *Oidipous epi Kolōnō*) the old, blind Oedipus has spent many years wandering in exile after being rejected by his sons and the city of Thebes. Oedipus has been cared for only by his daughters Antigone and Ismene. He arrives at a sacred grove at Colonus, a village close by Athens (and the home of Sophocles himself). There Oedipus is guaranteed protection by Theseus, the noble king of Athens. Theseus does indeed protect Oedipus from the importunate pleadings of his brother-in-law, Creon, for Oedipus to protect Thebes. Oedipus himself rejects the entreaties of his son Polyneices, who is bent on attacking Thebes and whom Oedipus solemnly curses. Finally Oedipus departs to a mysterious death; he is apparently swallowed into the earth of Colonus, where he will become a benevolent power and a mysterious source of defense to the land that has given him final refuge. The play is remarkable for the melancholy, beauty, and power of its lyric odes and for the spiritual and moral authority with which it invests the figure of Oedipus.

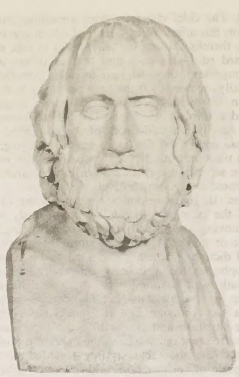
Trackers. Four hundred lines of this satyr play survive. The plot of *Trackers* (Greek *Ichneutai*) is based on two stories about the miraculous early deeds of the god Hermes: that the infant, growing to maturity in a few days, stole cattle from Apollo, baffling discovery by reversing the animals' hoof marks, and that he invented the lyre by fitting strings to a tortoise shell. In this play the trackers are the chorus of satyrs, who are looking for the cattle; they are amusingly dumbfounded at the sound of the new instrument Hermes has invented. Enough of the play survives to give an impression of its style; it is a genial, uncomplicated travesty of the tragic manner, and the antics of the chorus were apparently the chief source of amusement. (T.W./O.T.)

Satyr play

Euripides

LIFE AND CAREER

It is possible to reconstruct only the sketchiest biography of Euripides, the youngest of classical Athens' three great tragedians. Euripides' mother's name was Cleito; his father's name was Mnesarchus or Mnesarchides. One tradition states that his mother was a greengrocer who sold herbs in the marketplace. Aristophanes joked about this in



Euripides, marble herm copied from a Greek original, c. 340–330 BC. In the Museo Archeologico Nazionale, Naples.

By courtesy of the Soprintendenza alle Antichita della Campania, Naples

comedy after comedy; but there is better indirect evidence that Euripides came of a well-off family. He was probably born in 484. Euripides first received the honour of being chosen to compete in the dramatic festival in 455, and he won his first victory in 441. Euripides left Athens for good in 408, accepting an invitation from Archelaus, king of Macedonia. He died in Macedonia in 406.

Euripides' only known public activity was his service on a diplomatic mission to Syracuse in Sicily. He was passionately interested in ideas, however, and owned a large library. He is said to have associated with Protagoras, Anaxagoras, and other Sophists and philosopher-scientists. His acquaintance with new ideas brought him restlessness rather than conviction, however, and his questioning attitude toward traditional Greek religion is reflected in some of his plays. Of Euripides' private life, little can be said. Later tradition invented for him a spectacularly disastrous married life. It is known that he had a wife called Melito and produced three sons. One of these was something of a poet and produced the *Bacchants* after his father's death. He may also have completed his father's unfinished play *Iphigenia at Aulis.*

Uncertain biographical accounts

The ancients knew of 92 plays composed by Euripides. Nineteen plays are extant, if one of disputed authorship is included. At only four festivals was Euripides awarded the first prize—the fourth posthumously, for the tetralogy that included *Bacchants* and *Iphigenia at Aulis.* As Sophocles won perhaps as many as 24 victories, it is clear that Euripides was comparatively unsuccessful. More to the point is that on more than 20 occasions Euripides was chosen, out of all contestants, to be one of the three laureates of the year. Furthermore, the regularity with which Aristophanes parodied him is proof enough that Euripides' work commanded attention. It is often said that disappointment at his plays' reception in Athens was one of the reasons for his leaving his native city in his old age; but there are other reasons why an old poet might have left Athens in the 23rd year of the Peloponnesian War.

DRAMATIC AND LITERARY ACHIEVEMENTS

Euripides' plays exhibit his iconoclastic, rationalizing attitude toward both religious belief and the ancient legends and myths that formed the traditional subject matter for Greek drama. These legends seem to have been for him a mere collection of stories without any particular authority. He also apparently rejected the gods of Homeric theology, whom he frequently depicts as irrational, petulant, and singularly uninterested in meting out "divine justice."

That the gods are so often presented on the stage by Euripides is partly due to their convenience as a source of information that could not otherwise be made available to the audience.

Given this attitude of sophisticated doubt on his part, Euripides invents protagonists who are quite different from the larger-than-life characters drawn with such conviction by Aeschylus and Sophocles. They are, for the most part, commonplace, down-to-earth men and women who have all the flaws and vulnerabilities ordinarily associated with human beings. Furthermore, Euripides makes his characters express the doubts, the problems and controversies, and in general the ideas and feelings of his own time. They sometimes even take time off from the dramatic action to debate each other on matters of current philosophical or social interest.

Euripides differed from Aeschylus and Sophocles in making his characters' tragic fates stem almost entirely from their own flawed natures and uncontrolled passions. Chance, disorder, and human irrationality and immorality frequently result not in an eventual reconciliation or moral resolution but in apparently meaningless suffering that is looked upon with indifference by the gods. The power of this type of drama lies in the frightening and ghastly situations it creates and in the melodramatic, even sensational, emotional effects of its characters' tragic crises.

Psychological realism

Given this strong strain of psychological realism, Euripides shows moments of brilliant insight into his characters, especially in scenes of love and madness. His depictions of women deserve particular attention; it is easy to extract from his plays a long list of heroines who are fierce, treacherous, or adulterous, or all three at once. Misogyny is altogether too simple an explanation here, although Euripides' reputation in his own day was that of a woman hater, and a play by Aristophanes, *Women at the Thesmophoria,* comically depicts the indignation of the Athenian women at their portrayal by Euripides.

The chief structural peculiarities of Euripides' plays are his use of prologues and of the providential appearance of a god (deus ex machina) at the play's end. Almost all of the plays start with a monologue that is in effect a bare chronicle explaining the situation and characters with which the action begins. Similarly, the god's epilogue at the end of the play serves to reveal the future fortunes of the characters. This latter device has been criticized as clumsy or artificial by modern authorities, but it was presumably more palatable to the audiences of Euripides' own time. Another striking feature of his plays is that over time Euripides found less and less use for the chorus; in his successive works it tends to grow detached from the dramatic action.

The word habitually used in antiquity to describe Euripides' ordinary style of dramatic speech is *lalia* ("chatter"), alluding probably both to its comparatively light weight and to the volubility of his characters of all classes. Notwithstanding this, Euripides' lyrics at times have considerable charm and sweetness. In the works written after 415 BC his lyrics underwent a change, becoming more emotional and luxuriant. At its worst this style is hardly distinguishable from Aristophanes' parody of it in his comedy *Frogs,* but where frenzied emotion is appropriate, as in the tragedy *Bacchants,* Euripides' songs are unsurpassed in their power and beauty.

The tragicomedies

During the last decade of his career Euripides began to write "tragedies" that might actually be called romantic dramas, or tragicomedies with happy endings. These plays have a highly organized structure leading to a recognition scene in which the discovery of a character's true identity produces a complete change in the situation, and in general a happy one. Extant plays in this style include *Ion, Iphigenia Among the Taurians,* and *Helen.* Plays of the tragicomedy type seem to anticipate the New Comedy of the 4th century BC.

The fame and popularity of Euripides eclipsed that of Aeschylus and Sophocles in the cosmopolitan Hellenistic period. The austere, lofty, essentially political and "religious" tragedy of Aeschylus and Sophocles had less appeal than that of Euripides, with its more accessible realism and its obviously emotional, even sensational, effects. Euripides thus became the most popular of the three for revivals of his plays in later antiquity; this is probably why at least 18 of his plays have survived compared to seven each for Aeschylus and Sophocles, and why the extant fragmentary quotations from his works are more numerous than those of Aeschylus and Sophocles put together.

THE PLAYS

The dates of production of nine of Euripides' plays are known with some certainty from evidence that goes back to the official Athenian records. Those plays whose dates are prefixed by *c.* can be dated to within a few years by the internal evidence of Euripides' changing metrical techniques.

Alcestis. Though tragic in form, *Alcestis* (438 BC; Greek *Alkēstis*) ends happily and took the place of the satyr play that normally followed the three tragedies. King Admetus is doomed to die shortly, but he will be allowed a second life if he can find someone willing to die in his place. His wife, Alcestis, voluntarily dies in place of her husband, who sees too late that the fact and manner of her dying will blight his life. But Admetus' old friend Heracles shows up and rescues Alcestis from the clutches of Death, restoring her to her happy and relieved husband.

Medea. One of Euripides' most powerful and best known plays, *Medea* (431 BC; Greek *Mēdeia*) is a remarkable study of the mistreatment of a woman and of her ruthless revenge. The Colchian princess Medea has been taken by the hero Jason to be his wife. They have lived happily for some years at Corinth and have two sons. But then Jason casts Medea off and decides to marry the Princess of Corinth. Medea is determined on revenge, and after a dreadful mental struggle between her passionate sense of injury and her love for her children, she decides to punish her husband by murdering both the Corinthian princess and their own sons, thereby leaving her husband to grow old with neither wife nor child. She steels herself to commit these deeds and then escapes in the chariot of her grandfather, the sun-god Helios, leaving Jason without even the satisfaction of punishing her for her crimes. Euripides succeeds in evoking sympathy for the figure of Medea, who becomes to some extent a representative of women's oppression in general.

Children of Heracles. The plot of *Children of Heracles* (430 BC; Greek *Hērakleidai*) concerns the Athenians' defense of the young children of the dead Heracles from the murderous intentions of King Eurystheus of Argos. The play is basically a simple glorification of Athens.

Hippolytus. In *Hippolytus* (428 BC; Greek *Hippolytos*) Aphrodite, the goddess of love and sexual desire, destroys Hippolytus, a lover of outdoor sports who is repelled by sexual passion and who is instead devoted to the virgin huntress Artemis. Aphrodite makes Phaedra, wife of Theseus, the king of Athens, fall violently in love with her stepson Hippolytus. Phaedra is deeply ashamed of her illicit passion, but when Hippolytus angrily rejects her love she is so mortified by his denunciation that she cannot forbear from falsely accusing him of rape before she kills herself. Her accusation provokes Theseus into pronouncing a curse on his son that eventually leads to Hippolytus' death. But Artemis reveals Hippolytus' innocence before he dies, and the young man is able to forgive his father, thus freeing Theseus from the dreadful stain of bloodguilt. Given the nature of its plot, the play is remarkable for its propriety.

Andromache. This play is set in the aftermath of the Trojan War. After an exciting beginning marked by strong anti-Spartan feeling, most of the original characters in *Andromache* (*c.* 426 BC) disappear and the interest is dissipated.

Miseries of war

Hecuba. Also set in the aftermath of the Trojan War, *Hecuba* (*c.* 425 BC; Greek *Hekabē*) shows the double disaster that reduces the aged Trojan queen Hecuba, now a widowed slave, by sheer weight of hatred and misery to a mere animal ferocity. Hecuba first loses her daughter Polyxena, who is taken off to be sacrificed to the ghost of Achilles. Hecuba then discovers the corpse of her last son, Polydorus, who has been murdered by his Thracian host, Polymestor. Hecuba eventually persuades the Greek

commander Agamemnon to allow her to take vengeance; she and her women then blind Polymestor and murder his two young sons. Such is the power of misery to deprave, and the play's closing prophecy of Hecuba's future transformation into a bitch seems appropriate.

Suppliants. The title figures of *Suppliants* (*c.* 423 BC; Greek *Hiketides;* Latin *Supplices*) are the mothers of the Argive leaders who have been killed while attacking Thebes. The bodies of their sons have been left unburied by the Thebans, and they eventually persuade the Athenians to recover them. It is disputed whether the play is a straightforward eulogy of Athens and its democracy, or whether its sentiments are being expressed ironically.

Electra. The title character of *Electra* (*c.* 418 BC; Greek *Ēlektra*) and her brother Orestes murder their mother, Clytemnestra, in retribution for her murder of their father, Agamemnon. Electra herself is portrayed as a frustrated and resentful woman who finally lures her mother to her death by appealing to her maternal instincts. After the horrible murder both Electra and her reluctant accomplice Orestes are consumed by remorse. This is a bitterly realistic and antiheroic play that draws a disturbingly convincing portrait of both Electra's sufferings and her unattractive personality.

Madness of Heracles. The title character of *Madness of Heracles* (*c.* 416 BC; Greek *Hēraklēs mainomenos;* Latin *Hercules furens*) is temporarily driven mad by the goddess Hera and kills his wife and children. Subsequently Heracles recovers his reason and, after recovering from suicidal despair, is taken to spend an honourable retirement at Athens.

Trojan Women. The setting of *Trojan Women* (415 BC; Greek *Trōades*) is the time immediately after the taking of Troy, and the play treats the sufferings of the wives and children of the city's defeated leaders, in particular the old Trojan queen Hecuba and her children. Hecuba's daughter Cassandra is taken off to be the concubine of Agamemnon, and then her daughter-in-law Andromache is led off to be the slave of Neoptolemus. Andromache's son Astyanax is taken from her to be hurled to his death from the walls of Troy. Finally, as Troy goes up in flames, Hecuba and the other Trojan women are taken off to the ships to face slavery in Greece. This play is a famous and powerful indictment of the barbarous cruelties of war. It was first produced only months after the Athenians captured the city-state of Melos, butchering its men and reducing its women to slavery, and the *Trojan Women*'s mood may well have been influenced by the Athenians' atrocities and the Melians' fate, which are both mirrored in the play.

Plays with recognition scenes

Ion. This tragicomedy's sombre action is reversed by a recognition scene. In *Ion* (*c.* 413 BC), Creusa, the queen of Athens, is married to an immigrant king, Xuthus, but the couple do not have any children. Years before, the Queen was raped by the god Apollo but abandoned the subsequent child. The boy Ion has grown up as a temple slave at Delphi, where the play is set. When they meet, mother and son feel a strong affinity, but when the Delphic oracle says the boy is the son of Xuthus, the Queen in her despairing childlessness plots to kill the young stranger who threatens to take over her inheritance. At the last minute they recognize each other by means of the cradle Creusa had long ago left with her baby. The play has a superficially satisfactory ending, but its portrayal of human suffering and of divine carelessness and mendacity is tinged with darker feelings.

Iphigenia Among the Taurians. This is another tragicomedy, composed chiefly of a recognition scene followed by a clever escape. The title character of *Iphigenia Among the Taurians* (*c.* 413 BC; Greek *Iphigeneia en Taurois;* Latin *Iphigenia in Tauris*) has been saved by the goddess Artemis from sacrifice by her father and now serves the goddess' temple at Tauris in Thrace. Iphigenia's brother Orestes is captured by the local tyrant and is delivered to her for sacrifice. She recognizes him, however, and after some exciting mishaps they manage to escape from Tauris with the help of Athena.

Helen. In this frankly light play, Euripides deflates one of the best known "facts" of Greek mythology, that Helen ran off adulterously with Paris to Troy. In *Helen* (412 BC; Greek *Helenē*) only a phantom went with Paris to Troy, and the real Helen pines faithfully in Egypt. When Menelaus on his way home from Troy is shipwrecked in Egypt, he is baffled by the duplicate Helen until the evaporation of the phantom allows his reunion with the real one. The pair then escape from the King of Egypt, who is keen to marry Helen, by an amusing artifice.

Phoenician Women. This is a diverse, many-charactered play whose original version has been tampered with. *Phoenician Women* (*c.* 409 BC; Greek *Phoinissai*) is set at Thebes and concerns the mutual slaughter of the two sons of Oedipus, Eteocles and Polyneices.

Orestes. In this play Euripides makes nonsense of the old story of Orestes' murder of his mother, Clytemnestra, by setting the play in a world where courts of law already exist. In *Orestes* (408 BC), the main character, his sister Electra, and his cousin and friend Pylades are condemned to death by the men of Argos for the murder. Their uncle Menelaus is too spineless to defend them, and they are finally reduced to plotting to kill Menelaus' wife, Helen, and abduct her innocent daughter. This chaos of violence and attempted murder is only resolved by the deus ex machina Apollo, who appears and restores harmony at the end of the play.

Iphigenia at Aulis. The Greek fleet is becalmed at Aulis and is thus unable to convey the expeditionary force against Troy. Agamemnon learns that he must sacrifice his daughter Iphigenia as a means of appeasing the goddess Artemis, who has caused the unfavourable weather. Agamemnon lures his daughter into coming to Aulis to be sacrificed by pretending that she will marry Achilles. Once the truth is out, Iphigenia, after begging pathetically for her life, goes willingly to her death. Though incomplete and corrupted by later adapters, *Iphigenia at Aulis* (*c.* 406 BC; Greek *Iphigeneia en Aulidi*) is a fine tragedy whose realistic atmosphere is heightened by several subtle and poignant scenes between its main characters.

Bacchants. This play is regarded by many as Euripides' masterpiece. In *Bacchants* (*c.* 406 BC; Greek *Bakchai;* Latin *Bacchae*) the god Dionysus arrives in Greece from Asia intending to introduce his orgiastic worship there. He is disguised as a charismatic young Asian holy man and is accompanied by his women votaries, who make up the play's chorus. He expects to be accepted first in Thebes, but the Thebans reject his divinity and refuse to worship him, and the city's young king, Pentheus, tries to arrest him. In the end Dionysus drives Pentheus insane and leads him to the mountains, where Pentheus' own mother, Agave, and the women of Thebes in a bacchic frenzy tear him to pieces. Agave returns to Thebes triumphant carrying Pentheus' head, and her father, Cadmus, has to lead her back to sanity and recognition. The play shows how the liberating and ecstatic side of the Dionysiac religion must be balanced against the dangerous irresponsibility that goes with the Dionysiac loss of reason and self-consciousness.

Late masterpiece

Cyclops. *Cyclops* (Greek *Kyklōps*) is the only complete surviving satyr play. The play's cowardly, lazy satyrs with their disgraceful old father Silenus are slaves of the man-eating one-eyed Cyclops Polyphemus in Sicily. Odysseus arrives, driven to Sicily by adverse weather, and eventually succeeds (as in Homer's *Odyssey*) in blinding the Cyclops. He thus enables the Cyclops' victims to escape.

(H.D.F.K./O.T.)

Aristophanes

The most famous of all comic dramatists of ancient Greece, Aristophanes is also the one whose works have been preserved in the greatest quantity. He is the only extant representative of the Old Comedy, that is, of the phase of comic dramaturgy in which chorus, mime, and burlesque still played a considerable part and which was characterized by bold fantasy, merciless invective and outrageous satire, unabashedly licentious humour, and a marked freedom of political criticism. But Aristophanes belongs to the end of this phase, and, indeed, his last extant play, which has no choric element at all, may well

be regarded as the only extant specimen of the short-lived Middle Comedy, which, before the end of the 4th century BC, was to be superseded in turn by the milder and more realistic social satire of the New Comedy.

LIFE AND CAREER

Little is known about the life of Aristophanes, and most of the known facts are derived from references in his own plays. Born *c.* 450 BC, he was an Athenian citizen belonging to the *deme,* or clan, named Pandionis, but his actual birthplace is uncertain. (The fact that he or his father, Philippus, owned property on the island of Aegina may have been the cause of an accusation by his fellow citizens that he was not of Athenian birth.) He began his dramatic career in 427 BC with a play, the *Daitaleis* (*The Banqueters*), which appears, from surviving fragments, to have been a satire on his contemporaries' educational and moral theories. He is thought to have written about 40 plays in all. A large part of his work is concerned with the social, literary, and philosophical life of Athens itself and with themes provoked by the great Peloponnesian War (431–404 BC). This war was essentially a conflict between imperialist Athens and conservative Sparta and so was long the dominant issue in Athenian politics. Aristophanes was naturally an opponent of the more or less bellicose statesmen who controlled the government of Athens throughout the better part of his maturity. Aristophanes lived to see the revival of Athens after its defeat by Sparta. He died in about 388 BC.

Importance of Peloponnesian War on his work

DRAMATIC AND LITERARY ACHIEVEMENTS

Aristophanes' reputation has stood the test of time; his plays have been frequently produced on the 20th-century stage in numerous translations, which manage with varying degrees of success to convey the flavour of Aristophanes' puns, witticisms, and topical allusions. But it is not easy to say why his comedies still appeal to an audience almost 2,500 years after they were written. In the matter of plot construction Aristophanes' comedies are often loosely put together, are full of strangely inconsequential episodes, and often degenerate at their end into a series of disconnected and boisterous episodes. Aristophanes' greatness lies in the wittiness of his dialogue; in his generally good-humoured though occasionally malevolent satire; in the brilliance of his parody, especially when he mocks the controversial tragedian Euripides; in the ingenuity and inventiveness, not to say the laughable absurdity, of his comic scenes born of imaginative fantasy; in the peculiar charm of his choric songs, whose freshness can still be conveyed in languages other than Greek; and, at least for audiences of a permissive age, in the licentious frankness of many scenes and allusions in his comedies.

THE PLAYS

Babylonians. This comedy, which is extant only in fragments, was produced at the festival of the Great Dionysia. The festival was attended by delegates of the city-states, which were theoretically "allies" but were in practice satellites of Athens. Because *Babylonians* (426 BC; Greek *Babylōnioi*) not only virulently attacked Cleon, the demagogue then in power in Athens, but also showed the "allies" as the slaves of the Athenian Demos (a personification of the Athenian citizen electorate), Aristophanes was impeached by Cleon. Though the details are not known, he seems to have been let off lightly.

Acharnians. This is the earliest of the 11 comedies of Aristophanes that have survived intact. *Acharnians* (425 BC; Greek *Acharneis*) is a forthright attack on the folly of the war. Its farmer-hero, Dicaeopolis, is tired of the Peloponnesian War and therefore secures a private peace treaty with the Spartans for himself in spite of the violent opposition of a chorus of embittered and bellicose old charcoal burners of Acharnae. Dicaeopolis takes advantage of his private treaty to trade with the allies of the Spartans. The Athenian commander Lamachus tries to stop him, but by the end of the play Lamachus slumps wounded and dejected while Dicaeopolis enjoys a peacetime life of food, wine, and sex.

Knights. This play shows how little Aristophanes was affected by the prosecution he had incurred for *Babylonians. Knights* (424 BC; Greek *Hippeis*) consists of a violent attack on the same demagogue, Cleon, who is depicted as the favourite slave of the stupid and irascible Demos until he is, at last, ousted from his position of influence and authority by Agoracritus, a sausage seller who is even more scoundrelly and impudent than Cleon.

Clouds. This play (423 BC; Greek *Nephelai*) is an attack on "modern" education and morals as imparted and taught by the radical intellectuals known as the Sophists. The main victim of the play is the leading Athenian thinker and teacher Socrates, who is purposely (and unfairly) given many of the standard characteristics of the Sophists. In the play Socrates is consulted by an old rogue, Strepsiades ("Twisterson"), who wants to evade his debts. The instruction at Socrates' academy, the Phrontisterion ("Thinking Shop"), which consists of making a wrong argument sound right, enables Strepsiades' son to defend the beating of his own father. At the play's end the Phrontisterion is burned to the ground.

Attack on Sophists

Wasps. This comedy satirized the litigiousness of the Athenians in the person of the mean and waspish old man Philocleon ("Love-Cleon"), who has a passion for serving on juries. In *Wasps* (422 BC; Greek *Sphēkes*) Philocleon's son, Bdelycleon ("Loathe-Cleon"), arranges for his father to hold a "court" at home; but, since the first "case" to be heard is that of the house dog accused of the theft of a cheese, Philocleon is finally cured of his passion for the law courts and instead becomes a boastful and uproarious drunkard. The play's main political target is the exploitation by Cleon of the Athenian system of large subsidized juries.

Peace. This play was staged seven months or so after both Cleon and Brasidas, the two main champions of the war policy on the Athenian and Spartan sides respectively, had been killed in battle and, indeed, only a few weeks before the ratification of the Peace of Nicias (? March 421 BC), which suspended hostilities between Athens and Sparta for six uneasy years. In *Peace* (421 BC; Greek *Eirēnē*) the war-weary farmer Trygaeus ("Vintager") flies to heaven on a monstrous dung beetle to find the lost goddess Peace, only to discover that the God of War has buried Peace in a pit. With the help of a chorus of farmers Trygaeus rescues her, and the play ends with a joyful celebration of marriage and fertility.

Birds. This play can be regarded merely as a "comedy of fantasy," but some scholars see *Birds* (414 BC; Greek *Ornithes*) as a political satire on the imperialistic dreams that had led the Athenians to undertake their ill-fated expedition of 415 BC to conquer Syracuse in Sicily. Peisthetaerus ("Trusty") is so disgusted with his city's bureaucracy that he persuades the birds to join him in building a new city that will be suspended in between heaven and earth; it is named Nephelokokkygia and is the original Cloudcuckooland. The city is built, and Peisthetaerus and his bird comrades must then fend off the undesirable humans who want to join them in their new Utopia. He and the birds finally even starve the Olympian gods into cooperating with them. *Birds* is Aristophanes' most fantastical play, but its escapist mood possibly echoes the dramatist's sense of Athens' impending decline.

Lysistrata. This comedy was written not long after the catastrophic defeat of the Athenian expedition to Sicily (413 BC) and not long before the revolt of the Four Hundred in Athens, whereby an oligarchic regime ready to make peace with Sparta was set up (411 BC). *Lysistrata* (411 BC; Greek *Lysistratē*) depicts the seizure of the Acropolis and of the treasury of Athens by the city's women who, at Lysistrata's instigation, have, together with all the women of Greece, declared a sex strike until such time as the men will make peace. The women defy their menfolk until the peace is arranged, after which both the Athenian and Spartan wives are reunited with their husbands. The play is a strange mixture of humour, indecency, gravity, and farce.

Plot of *Lysistrata*

Women at the Thesmophoria. In *Women at the Thesmophoria* (411 BC; Greek *Thesmophoriazousai*) Euripides has discovered that the women of Athens, angered by his constant attacks upon them in his tragedies, mean to

discuss during their coming festival (the Thesmophoria) the question of contriving his death. Euripides tries to persuade the effeminate Agathon, a tragic poet, to plead his cause. Agathon refuses, and Euripides persuades his brother-in-law Mnesilochus to undertake the assignment. Mnesilochus is disguised with great thoroughness as a woman and sent on his mission, but his true sex is discovered and he is at once seized by the women. There follow three scenes in which he tries unsuccessfully to escape; all three involve brilliant parodies of Euripides' tragedies, and all three attempts fail. Finally, Euripides himself arrives and succeeds in rescuing his advocate by promising never again to revile women.

Frogs. This is a literary comedy. In *Frogs* (405 BC; Greek *Batrachoi*) Dionysus, the god of drama, is concerned about the poor quality of present-day tragedy in Athens now that his recent favourite, Euripides, is dead. Dionysus disguises himself as the hero Heracles and goes down to Hades to bring Euripides back to the land of the living. As the result, however, of a competition arranged between Euripides and his great predecessor, Aeschylus, Dionysus is won over to the latter's cause and returns to earth with Aeschylus, instead, as the one more likely to help Athens in its troubles.

Women at the Ecclesia. In *Women at the Ecclesia* (*c.* 392 BC; Greek *Ekklēsiazousai*) the women of Athens dress up as men, take over the Ecclesia (the Athenian democratic assembly), and introduce a communistic system of wealth, sex, and property. It is not one of Aristophanes' more appealing plays.

Wealth. The last of the author's plays to be performed in his lifetime, *Wealth* (388 BC; Greek *Ploutos*) is a somewhat moralizing work and does not enhance his reputation—though, as suggested, it may have inaugurated the Middle Comedy.

Shortly after producing his *Wealth,* Aristophanes died, leaving two plays (now lost), the *Aiolosikon* and the *Kokalos,* which his son staged *c.* 387 BC; both of them are generally assumed to have been mythological burlesques.

(M.Pl./O.T.)

MAJOR WORKS

Aeschylus

WORKS: *Persai* (472 BC; *Persians*); *Hepta epi Thēbais* (467 BC; *Seven Against Thebes*); *Hiketides* (*c.* 463 BC; Latin trans., *Supplices;* Eng. trans., *Suppliants*); the trilogy known as the *Oresteia* (458 BC), comprising *Agamemnōn, Choēphoroi* (*Libation Bearers*), and *Eumenides*; *Promētheus desmōtēs* (date uncertain, probably late; *Prometheus Bound*).

TEXTS: The Greek text is available in Denys Page (ed.), *Aeschyli septem quae supersunt tragoedias* (1972, reprinted 1975), in the Oxford Classical Text series. Herbert Weir Smyth (trans.), *Aeschylus,* 2 vol. (1922–26, reprinted 1973–83), in the Loeb Classical Library series, presents the original text with an English translation.

RECOMMENDED EDITIONS: *Aeschylus,* 2 vol. (1953–56, reissued 1967), contains English translations of all the plays and is part of The Complete Greek Tragedies series, ed. by Richmond Lattimore and David Grene. Recent editions of the trilogy are *Oresteia,* trans. by Hugh Lloyd-Jones, 3 vol. (1979, reissued in 1 vol., 1982); and *The Oresteia,* trans. by Robert Fagles (1975, reissued 1984).

Sophocles

WORKS: *Aias mastigophoros* (probably before 441 BC; *Ajax*); *Antigonē* (*c.* 442–441 BC); *Oidipous Tyrannos* (soon after 430 BC; Latin trans., *Oedipus Rex;* Eng. trans., *Oedipus the King*); *Trachiniai* (possibly after 430 BC; *Trachinian Women*); *Elektra* (between 418 and 410 BC; *Electra*); *Philoktētēs* (409 BC; *Philoctetes*); *Oidipous epi Kolōnō* (produced posthumously in 401 BC; *Oedipus at Colonus*); *Ichneutai* (sizable fragments; date uncertain; *Trackers*).

TEXTS: The Greek text is available in A.C. Pearson (ed.), *Sophoclis fabulae* (1924, reprinted 1975), in the Oxford Classical Text series; and R.D. Dawe (ed.), *Sophoclis tragoediae,* vol. 1 and 2 (1975–79), part of the Teubner series. Greek text with English translation is presented in R.C. Jebb (ed. and trans.), *Sophocles,* 7 vol. (1883–96, reprinted from various editions, 1976).

RECOMMENDED EDITIONS: *Sophocles,* 2 vol. (1954–69), contains translations of all the plays and is part of The Complete Greek Tragedies series, ed. by Richmond Lattimore and David Grene. See also Robert Fagles (trans.), *The Three Theban Plays* (1982, reissued 1984), which includes *Antigonē, Oedipus the King,* and *Oedipus at Colonus.*

Euripides

WORKS: *Alkēstis* (438 BC; *Alcestis*); *Mēdeia* (431 BC; *Medea*); *Hērakleidai* (*c.* 430 BC; *Children of Heracles*); *Hippolytos* (428 BC; *Hippolytus*); *Andromachē* (*c.* 426 BC); *Hekabē* (*c.* 425 BC; *Hecuba*); *Hiketides* (*c.* 423 BC; Latin trans., *Supplices;* Eng. trans., *Suppliants*); *Elektra* (*c.* 418 BC; *Electra*); *Hēraklēs mainomenos* (*c.* 416 BC; Latin trans., *Hercules furens;* Eng. trans., *Madness of Heracles*); *Trōades* (415 BC; *Trojan Women*); *Iōn* (*c.* 413 BC); *Iphigeneia en Taurois* (*c.* 413 BC; Latin trans., *Iphigenia in Tauris;* Eng. trans., *Iphigenia Among the Taurians*); *Helenē* (412 BC; *Helen*); *Phoinissai* (*c.* 409 BC; *Phoenician Women*); *Orestēs* (408 BC); *Iphigeneia en Aulidi* (*c.* 406 BC; *Iphigenia at Aulis*); *Bakchai* (*c.* 406 BC; Latin trans., *Bacchae;* Eng. trans., *Bacchants*); *Kyklōps* (date unknown; *Cyclops*); *Rhēsos* (authorship disputed; date unknown; *Rhesus*); *Hypsipylē* (sizable fragments; date uncertain).

TEXTS: The Greek text is available in *Euripidis fabulae,* ed. by J. Diggle (1981–), a new Oxford Classical Text replacing the 3-vol. text ed. by Gilbert Murray, 1902–09; 2 vol. of the Diggle work have appeared to 1986.

RECOMMENDED EDITIONS: *Euripides,* 5 vol. (1955–59), contains English translations of all the plays and is part of The Complete Greek Tragedies series, ed. by Richmond Lattimore and David Grene. All the complete plays excluding *Cyclops* and *Rhesus* are found in the Penguin Classics series, trans. by Philip Vellacott, *Three Plays* (1953, reissued 1974), *Medea, and Other Plays* (1963), *Orestes, and Other Plays* (1972), and *The Bacchae, and Other Plays,* rev. ed. (1972). See also Geoffrey S. Kirk (trans.), *The Bacchae* (1970, reprinted 1979 as *The Bacchae of Euripides*).

Aristophanes

WORKS: *Babylōnioi* (426 BC; *Babylonians*); *Acharneis* (425 BC; *Acharnians*); *Hippeis* (424 BC; *Knights*); *Nephelai* (423 BC; *Clouds*); *Sphēkes* (422 BC; *Wasps*); *Eirēnē* (421 BC; *Peace*); *Ornithes* (414 BC; *Birds*); *Lysistratē* (411 BC; *Lysistrata*); *Thesmophoriazousai* (411 BC; *Women at the Thesmophoria*); *Batrachoi* (405 BC; *Frogs*); *Ekklēsiazousai* (*c.* 392 BC; *Women at the Ecclesia*); *Ploutos* (388 BC; *Wealth*).

TEXTS: The Greek text is available in Victor Coulon (ed.) and Hilaire Van Daele (trans.), *Aristophane,* 5 vol. (1923–30), part of the Budé series. Greek text with English translation is presented in Benjamin Bickley Rogers (ed. and trans.), *Aristophanous komodiai: The Comedies of Aristophanes,* 6 vol. in 7 (1902–16; reissued as part of the Loeb Classical Library series, 3 vol., 1924, reissued 1979–82); and in Alan H. Sommerstein (ed. and trans.), *The Comedies of Aristophanes* (1980–), five plays in single volumes having appeared to 1986.

RECOMMENDED EDITIONS: All the plays appear in English translation in three separate volumes in the Penguin Classics series, trans. respectively by David Barrett (1964, reprinted 1976), by Alan H. Sommerstein (1973, reprinted 1977), and by David Barrett and Alan H. Sommerstein (1978).

BIBLIOGRAPHY

Greek tragedy in general: H.D.F. KITTO, *Greek Tragedy: A Literary Study,* 3rd ed. (1961, reissued 1976), a lively survey, but becoming dated; RICHMOND LATTIMORE, *Story Patterns in Greek Tragedy* (1964, reissued 1969); ARTHUR PICKARD-CAMBRIDGE, *The Dramatic Festivals of Athens,* 2nd ed. rev. by JOHN GOULD and D.M. LEWIS (1968), a standard work on the practical arrangements; H.C. BALDRY, *The Greek Tragic Theatre* (1971), a simple, orthodox introduction; ALBIN LESKY, *Greek Tragedy* (1983; originally published in German, 3rd rev. ed., 1972); ERIKA SIMON, *The Ancient Theatre* (1982; originally published in German, 2nd ed., 1981), a concise and expert introduction to staging; JEAN-PIERRE VERNANT and PIERRE VIDAL-NAQUET, *Tragedy and Myth in Ancient Greece* (1981; originally published in French, 1972), stimulating structuralist essays; BRIAN VICKERS, *Towards Greek Tragedy: Drama, Myth, Society* (1973, reprinted 1979), long but thought-provoking; OLIVER TAPLIN, *Greek Tragedy in Action* (1978), emphasis on the significance of performance; BERNARD KNOX, *Word and Action: Essays on the Ancient Theater* (1979), a collection of important essays; DONALD J. MASTRONARDE, *Contact and Discontinuity: Some Conventions of Speech and Action on the Greek Tragic Stage* (1979), a specialist study of dialogue; R.G.A. BUXTON, *Persuasion in Greek Tragedy: A Study of Peitho* (1982); and ERICH SEGAL (ed.), *Greek Tragedy: Modern Essays in Criticism* (U.K. title, *Oxford Readings in Greek Tragedy,* 1983), a well-chosen and varied selection of articles.

Aeschylus: Studies of special topics relating to Aeschylus' plays include ANTHONY J. PODLECKI, *The Political Background of Aeschylean Tragedy* (1966); A.F. GARVIE, *Aeschylus' "Supplices": Play and Trilogy* (1969); GEORGE THOMSON, *Aeschylus and Athens: A Study in the Social Origins of Drama,* 4th ed. (1973), a Marxist study; ANNE LEBECK, *The Oresteia: A Study in Language and Structure* (1971), on the significant connections of imagery; R.P. WINNINGTON-INGRAM, *Studies in*

Aeschylus (1983), a collection of insightful essays; THOMAS G. ROSENMEYER, *The Art of Aeschylus* (1982), a critical study; D.J. CONACHER, *Aeschylus' "Prometheus Bound": A Literary Commentary* (1980); MARK GRIFFITH, *The Authenticity of "Prometheus Bound"* (1977), a powerful attack on authenticity; and OLIVER TAPLIN, *The Stagecraft of Aeschylus: The Dramatic Use of Exits and Entrances in Greek Tragedy* (1978), on dramatic techniques and meanings.

Sophocles: Influential interpretations of Sophocles' life and works are T.B.L. WEBSTER, *An Introduction to Sophocles*, 2nd ed. (1969, reprinted 1979); G.H. GELLIE, *Sophocles: A Reading* (1972); R.P. WINNINGTON-INGRAM, *Sophocles: An Interpretation* (1980), an important study; CHARLES SEGAL, *Tragedy and Civilization: An Interpretation of Sophocles* (1981), a structuralist approach; BERNARD M.W. KNOX, *The Heroic Temper: Studies in Sophoclean Tragedy* (1964, reprinted 1983), an important study of Sophocles' stubborn "heroes"; KARL REINHARDT, *Sophocles* (1979; originally published in German, 3rd ed., 1947), a great classic; G.M. KIRKWOOD, *A Study of Sophoclean Drama* (1958, reprinted 1967); R.W.B. BURTON, *The Chorus in Sophocles' Tragedies* (1980); and DAVID SEALE, *Vision and Stagecraft in Sophocles* (1982).

Euripides: Critical works include G.M.A. GRUBE, *The Drama of Euripides* (1941, reprinted 1973), a survey; D.J. CONACHER, *Euripidean Drama: Myth, Theme and Structure* (1967), a helpful play-by-play survey, including a useful bibliography; GILBERT MURRAY, *Euripides and His Age*, 2nd ed. (1946, reissued 1979), an idealistic introduction; T.B.L. WEBSTER, *The Tragedies of Euripides* (1967), helpful on the lost tragedies; R.P. WINNINGTON-INGRAM, *Euripides and Dionysus: An Interpretation of the Bacchae* (1948, reprinted 1969), an enterprising study; WILLIAM RITCHIE, *The Authenticity of the Rhesus of Euripides* (1964), a scholarly though unsuccessful case for authorship; ANNE PIPPIN BURNETT, *Catastrophe Survived: Euripides Plays of Mixed Reversal* (1971, reprinted 1985), an original interpretation of seven plays, PIETRO PUCCI, *The Violence of Pity in Euripides' "Medea"* (1980); SHIRLEY A. BARLOW, *The Imagery of Euripides: A Study in the Dramatic Use of Pictorial Language* (1971, reprinted 1974); and HELENE P. FOLEY, *Ritual Irony: Poetry and Sacrifices in Euripides* (1985), a structuralist study of four plays.

Aristophanes: Critical studies include GILBERT MURRAY, *Aristophanes* (1933, reprinted 1965), still lively; VICTOR EHRENBERG, *The People of Aristophanes: A Sociology of Old Attic Comedy*, 2nd ed. (1951, reprinted with revisions 1974); CEDRIC H. WHITMAN, *Aristophanes and the Comic Hero* (1964), concentrating on the central figure of each play; F.H. SANDBACH, *The Comic Theatre of Greece and Rome* (1977), especially the first chapters on Old Comedy; CARROLL MOULTON, *Aristophanic Poetry* (1981), artistic unity through poetry and imagery; ROSEMARY M. HARRIOTT, *Aristophanes: Poet & Dramatist* (1986), a study of sophistication of literary technique; K.J. DOVER, *Aristophanic Comedy* (1972), an introduction; KENNETH MCLEISH, *The Theatre of Aristophanes* (1980), on the mechanics of raising laughs; and JEFFREY HENDERSON (ed.), *Aristophanes: Essays in Interpretation* (1980).

(A.J.P./T.W./H.D.F.K./M.Pl./O.T.)

Greek Literature

Greek literature has a continuous history extending from the 1st millennium BC to the present day. From the beginning its writers were Greeks living not only in Greece proper but also in Asia Minor, the Aegean Islands, and Magna Graecia (Sicily and southern Italy). Later, after the conquests of Alexander the Great, Greek became the common language of the eastern Mediterranean lands and then of the Byzantine Empire. Literature in Greek was produced not only over a much wider area but also by those whose mother tongue was not Greek. Even before the Turkish conquest (1453) the area had begun to shrink again, and now it is chiefly confined to the territory of the republic of Greece.

This article is divided into the following sections:

Ancient Greek literature 352
- Stylistic periods 352
  - Preclassical period, to the end of the 6th century BC
  - Classical period, 5th and 4th centuries BC
  - Hellenistic and Greco-Roman periods
- The genres 353
  - Epic narrative
  - Lyric poetry
  - Tragedy
  - Comedy
  - History
  - Rhetoric and oratory
  - Philosophical prose
  - Late forms of poetry
  - Late forms of prose

Byzantine literature 357
- General characteristics 357
- Principal forms of writing 357
  - Nonliturgical poetry
  - Liturgical poetry
  - Historical works
  - Rhetoric

Modern Greek literature (after 1453) 358
- Post-Byzantine period 358
- Independence and after 359
  - Old Athenian School
  - Heptanesian School
  - Demoticism and folklorism, 1880–1922
  - Literature from 1922

Bibliography 360

Ancient Greek literature

Of the literature of ancient Greece only a relatively small proportion survives. Yet it remains important, not only because much of it is of supreme quality but also because until the mid-19th century the greater part of the literature of the Western world was produced by writers who were familiar with the Greek tradition, either directly or through the medium of Latin, who were conscious that the forms they used were mostly of Greek invention, and who took for granted in their readers some familiarity with classical literature.

STYLISTIC PERIODS

The history of ancient Greek literature may be divided into three periods: preclassical (to the end of the 6th century BC); classical (5th and 4th centuries BC); and Hellenistic and Greco-Roman (3rd century BC onward).

Preclassical period, to the end of the 6th century BC. The Greeks created poetry before they made use of writing for literary purposes, and from the beginning their poetry was intended to be sung or recited. (The art of writing was little known before the 7th century BC. The script used in Crete and Mycenae [Linear B] is not known to have been employed for other than administrative purposes, and after the destruction of the Mycenaean cities it was forgotten.)

Its subject was myth—part legend, based sometimes on the dim memory of historical events; part folktale; and part primitive religious speculation. But since the myths were not associated with any religious dogma, even though they often treated of gods and heroic mortals, they were not authoritative and could be varied by a poet to express new concepts.

Thus, at an early stage Greek thought was advanced as poets refashioned their materials; and to this stage of preclassical literature belonged the epics ascribed to Homer, the *Iliad* and the *Odyssey,* retelling intermingled history and myth of the Mycenaean Age. These two great poems, standing at the beginning of Greek literature, established most of the literary conventions of the epic poem. The didactic poetry of Hesiod (*c.* 700 BC) was probably later in composition than Homer's epics and, though different in theme and treatment, continued the epic tradition.

Types of early Greek poetry

The several types of Greek lyric poetry originated in the preclassical period among the poets of the Aegean Islands and of Ionia on the coast of Asia Minor. Archilochus of Paros, of the 7th century BC, was the earliest Greek poet to employ the forms of elegy (in which the epic verse line alternated with a shorter line) and of personal lyric poetry. His work was very highly rated by the ancient Greeks but survives only in fragments; its forms and metrical patterns—the elegiac couplet and a variety of lyric metres—were taken up by a succession of Ionian poets. At the beginning of the 6th century Alcaeus and Sappho, composing in the Aeolic dialect of Lesbos, produced lyric poetry mostly in the metres named after them (the alcaic and the sapphic), which Horace was later to adapt to Latin poetry. No other poets of ancient Greece entered into so close a personal relationship with the reader as Alcaeus, Sappho, and Archilochus do. They were succeeded by Anacreon of Teos, in Ionia, who, like Archilochus, composed his lyrics in the Ionic dialect. Choral lyric, with musical accompaniment, belonged to the Dorian tradition and its dialect, and its representative poets in the period were Alcman in Sparta and Stesichorus in Sicily.

Both tragedy and comedy had their origins in Greece. "Tragic" choruses are said to have existed in Dorian Greece around 600 BC, and in a rudimentary dramatic form tragedy became part of the most famous of the Dionysian festivals, the Great, or City, Dionysia at Athens, about 534. Comedy, too, originated partly in Dorian Greece and developed in Attica, where it was officially recognized rather later than tragedy. Both were connected with the worship of Dionysus, god of fruitfulness and of wine and ecstasy.

Written codes of law were the earliest form of prose and were appearing by the end of the 7th century, when knowledge of reading and writing was becoming more widespread. No prose writer is known earlier than Pherecydes (*c.* 550 BC) of Syros, who wrote about the beginnings of the world: but the earliest considerable author was Hecataeus of Miletus, who wrote about both the mythical past and the geography of the Mediterranean and surrounding lands. To Aesop, a semi-historical, semi-mythological character of the mid-6th century, have been attributed the moralizing beast fables copied by later writers.

Classical period, 5th and 4th centuries BC. True tragedy was created by Aeschylus and reached its culmination with Sophocles and Euripides in the second half of the 5th century. Aristophanes, the greatest of the comedic poets, lived on into the 4th century, but the Old Comedy did not survive the fall of Athens in 404. (These four dramatists and their works are treated in detail in GREEK DRAMATISTS, THE CLASSICAL.)

The sublime themes of Aeschylean tragedy, in which man stands answerable to the gods and receives awe-inspiring

insight into their purposes, are exemplified in the three plays of the *Oresteia*. The tragedy of Sophocles made a progress toward both dramatic complexity and naturalness while remaining orthodox in its treatment of religious and moral issues. Euripides handled his themes on the plane of skeptical enlightenment and doubted the traditional picture of the gods. Corresponding development of dramatic realization accompanied the shift of vision: the number of individual actors was raised to three, each capable of taking several parts.

The three stages of dramatic comedy

The Old Comedy of Aristophanes was established later than tragedy but preserved more obvious traces of its origin in ritual; for the vigour, wit, and indecency with which it keenly satirized public issues and prominent persons clearly derived from the primitive ribaldry of the Dionysian festival. Aristophanes' last comedies show a transition, indicated by the dwindling importance of the chorus, toward the Middle Comedy, of which no plays are extant. This phase was followed toward the beginning of the 3rd century by the New Comedy, introduced by Menander, which turned for its subjects to the private fictional world of ordinary people. Later adaptations of New Comedy in Latin by Plautus and Terence carried the influence of his work on to medieval and modern times.

In the 5th century, Pindar, the greatest of the Greek choral lyrists, stood outside the main Ionic-Attic stream and embodied in his splendid odes a vision of the world seen in terms of aristocratic values that were already growing obsolete. Greek prose came to maturity in this period. Earlier writers such as Anaxagoras the philosopher and Protagoras the Sophist used the traditional Ionic dialect, as did Herodotus the historian. His successors in history, Thucydides and Xenophon, wrote in Attic.

The works of Plato and Aristotle, of the 4th century, are the most important of all the products of Greek culture in the intellectual history of the West. They were preoccupied with ethics, metaphysics, and politics as man's highest study and, in the case of Aristotle, extended the range to include physics, natural history, psychology, and literary criticism. They have formed the basis of Western philosophy and, indeed, they determined, for centuries to come, the development of European thought. (For detailed treatment see ARISTOTELIANISM, ARISTOTLE AND and PLATONISM, PLATO AND.)

This was also a golden age for rhetoric and oratory, first taught by Corax of Syracuse in the 5th century. The study of rhetoric and oratory raised questions of truth and morality in argument, and thus it was of concern to the philosopher as well as to the advocate and the politician and was expounded by teachers, among whom Isocrates was outstanding. The orations of Demosthenes, a statesman of 4th-century Athens and the most famous of Greek orators, are preeminent for force and power.

Hellenistic and Greco-Roman periods. In the huge empire of Alexander the Great, Macedonians and Greeks composed the new governing class; and Greek became the language of administration, a new composite dialect based to some extent on Attic and called the Koine, or common language. Everywhere the traditional city-state was in decline, and the individual was becoming aware of his isolation and seeking consolidation and satisfaction outside corporate society. Artistic creation now came under private patronage, and, except for Athenian comedy, compositions were intended for a small, select audience that admired polish, erudition, and subtlety.

An event of great importance for the development of new tendencies was the founding of the Museum, the shrine of the Muses with its enormous library, at Alexandria. The chief librarian was sometimes a poet as well as tutor of the heir apparent. The task of accumulating and preserving knowledge begun by the Sophists and continued by Aristotle and his adherents was for the first time properly endowed. Through the researches of the Alexandrian scholars, texts of ancient authors were preserved.

Importance of the Museum at Alexandria

The Hellenistic period lasted from the end of the 4th to the end of the 1st century BC. For the next three centuries, until Constantinople became the capital of the Byzantine Empire, Greek writers were conscious of belonging to a world of which Rome was the centre.

THE GENRES

Epic narrative. At the beginning of Greek literature stand the two great epics, the *Iliad* and the *Odyssey*. Some features of the poems reach far into the Mycenaean age, perhaps to 1500 BC, but the written works are traditionally ascribed to Homer; in something like their present form they probably date to the 8th century. (Homer and the works attributed to him are treated in detail in HOMERIC EPICS, THE.)

The *Iliad* and the *Odyssey* and the *Odyssey*

The *Iliad* and the *Odyssey* are primary examples of the epic narrative, which in antiquity was a long narrative poem, in an elevated style, celebrating heroic achievement. The *Iliad* is the tragic story of the wrath of Achilles, son of a goddess and richly endowed with all the qualities that make men admirable. With his readiness to sacrifice all to honour, Achilles embodies the Greek heroic ideal; and the contrast between his superb qualities and his short and troubled life reflects the sense of tragedy always prevalent in Greek thought. Whereas the *Iliad* is tragedy, the *Odyssey* is tragicomedy. It is an enriched version of the old folktale of the wanderer's return and of his triumph over those who were usurping his rights and importuning his wife at home. Odysseus too represents a Greek ideal. Though by no means inadequate in battle, he works mainly by craft and guile; and it is by mental superiority that he survives and prevails.

Both poems were based on plots that grip the reader, and the story is told in language that is simple and direct, yet eloquent. The *Iliad* and the *Odyssey*, though they are the oldest European poetry, are by no means primitive. They marked the fulfillment rather than the beginning of the literary form to which they belong. They were essentially oral poems, handed down, developed, and added to over a vast period of time, a theme upon which successive nameless poets freely improvised. The world they reflect is full of inconsistencies; weapons belong to both the Bronze and Iron Ages, and objects of the Mycenaean period jostle others from a time five centuries later. Certain mysteries remain: the date of the great poet or poets who gave structure and shape to the two epics; the social function of poems that take several days to recite; and the manner in which these poems came to be recorded in writing probably in the course of the 6th century BC.

In the ancient world the *Iliad* and the *Odyssey* stood in a class apart among epic poems. Of these, there were a large number known later as the epic cycle. They covered the whole story of the wars of Thebes and Troy as well as other famous myths. A number of shorter poems in epic style, the Homeric Hymns, are of considerable beauty.

Didactic poetry was not regarded by the Greeks as a form distinct from epic. Yet the poet Hesiod belonged to an altogether different world from Homer. He lived in Boeotia in central Greece about 700 BC. In his *Works and Days* he described the ways of peasant life and incidentally described the dreary Boeotian plain afflicted by heat, cold, and the oppression of a "gift-devouring" aristocracy. He believed passionately in a Zeus who cared about right and wrong and in Justice as Zeus's daughter. Hesiod's other surviving poem, the *Theogony*, attempts a systematic genealogy of the gods and recounts many myths associated with their part in the creation of the universe. Middle Eastern influence is clearly to be seen, especially in some of the cruder speculation about the origin of the universe.

Didactic poetry

By the end of the 6th century the epic tradition was a spent force until its revival in the Hellenistic period, and the few composers of epic narrative left little but their names.

Lyric poetry. Hesiod, unlike Homer, told something of himself, and the same is true of the lyric poets. Except for Pindar and Bacchylides at the end of the classical period, only fragments of the works of these poets survive. There had always been lyric poetry in Greece. All the great events of life as well as many occupations had their proper songs, and here too the way was open to advance from the anonymous to the individual poet.

The word "lyric" covers many sorts of poem. On the one hand, poems sung by individuals or chorus to the lyre, or sometimes to the flute, were called melic; elegiacs, in which the epic hexameter, or verse line of six metri-

cal feet, alternated with a shorter line, were traditionally associated with lamentation and a flute accompaniment; but they were also used for personal poetry, spoken as well as sung. Iambics (verse of iambs, or metrical units, basically of four alternately short and long syllables) were the verse form of the lampoon. Usually of an abusive or satirical—burlesque and parodying—character, they were not normally sung.

If Archilochus of Paros in fact was writing as early as 700 BC, he was the first of the post-epic poets. The fragments reflect the turbulent life of an embittered adventurer. Scorn both of men and of convention is the emotion that seems uppermost, and Archilochus was possessed of tremendous powers of invective. Of lesser stature than Archilochus were his successors, Semonides (often mistakenly identified with Simonides) of Amorgos and Hipponax of Ephesus.

Elegiac poets

Like the iambic writers, the elegiac poets came mostly from the islands and the Ionian regions of Asia Minor. Chief among them were Callinus of Ephesus and Mimnermus of Colophon. On the mainland of Greece, Tyrtaeus roused the spirit of the Spartans in their desperate struggle with the Messenian rebels in the years after 650. His martial poems are perhaps of more historical than literary interest. The same is to some extent true of the poems in elegiac, iambic, and trochaic (the latter a metre basically of four alternately long and short syllables) metres by Solon, an Athenian statesman, who used his poetry as a vehicle for propaganda. Xenophanes (born about 560 BC) rather in the same way used his poems to propagate his revolutionary religious and ethical ideas. The elegiacs attributed to Theognis seem to be poems of various dates suitable for use at drinking parties. Many of them were actually by Theognis himself (about 540 BC). Some give uninhibited expression to his hatred of the lower class rulers who had ousted the aristocracy of Megara; others are love poems to the boy Cyrnus; still others are gnomic commonplaces of Greek wisdom and morality.

About the beginning of the 6th century a new kind of poetry made its appearance in the island of Lesbos. It was composed in the local Aeolic dialect by members of the turbulent and factious aristocracy. Alcaeus (born about 620 BC), absorbed in political feuds and in civil war, expressed with striking directness searing hate and blind exultation. With the same directness and infinite grace, Sappho, a younger contemporary who seems to have enjoyed a freedom unknown to the women of mainland Greece, told of her loves and hates, though little is known of her relations with the girls named in her poems. The surviving works by their successor in personal lyric, Anacreon of Teos, suggest a more convivial amorousness.

Choral lyric was associated with the Dorian parts of the Greek mainland and the settlements in Sicily and south Italy, whereas poetry for solo performance was a product of the Ionian coast and the Aegean Islands. Thus choral song came to be conventionally written in a Doric dialect.

Choral lyric, which had lyre and flute accompaniments, was highly complicated in structure. It did not use traditional lines or stanzas; but the metre was formed afresh for each poem and never used again in exactly the same form, though the metrical units from which the stanzas, or strophes, were built up were drawn from a common stock and the form of the strophe was usually related to the accompanying dance. This elaborate art form was connected mainly with the cult of the gods or, as in the case of Pindar, the celebration of the victors in the great Hellenic games.

The earliest poet of choral lyrics of whose work anything has survived was Alcman of Sparta (about 620 BC). Somewhat later Stesichorus worked in Sicily, and his lyric versions of the great myths marked an important stage in the development of these stories. Simonides of Ceos, in Ionia, was among the most versatile of Greek poets. He was famed for his pathos, but today he is best known for his elegiac epitaphs, especially those on the Greek soldiers who fell in the struggle against Persia.

Pindar's odes

The supreme poet of choral lyric was Pindar from Thebes in Boeotia (born 518 or possibly 522–died after 446 BC), who is known mainly by his odes in honour of the victors at the great games held at Olympia, Delphi, the Isthmus of Corinth, and Nemea. The last of the lyric poets was Bacchylides (flourished 5th century BC), whose work, though often exquisite, is empty, reflecting the declining significance of myth.

Tragedy. Tragedy may have developed from the dithyramb, the choral cult song of the god Dionysus. Arion of Lesbos, who is said to have worked at Corinth in about 600, is credited with being the first to write serious poetry in this medium. Thespis (6th century BC), possibly combining with dithyrambs something of the Attic ritual of Dionysus of Eleutherae, is credited with having invented tragedy by introducing an actor who conversed with the leader of the chorus. These performances became a regular feature of the great festival of Dionysus at Athens about 534 BC. Aeschylus introduced a second actor, though his drama was still centred in the chorus, to whom, rather than to each other, his actors directed themselves.

The tragedies of Aeschylus, Sophocles, and Euripides

At the tragic contests at the Dionysia each of three competing poets produced three tragedies and a satyr play, or burlesque, in which there was a chorus of satyrs. Aeschylus, unlike later poets, usually made of his three tragedies a dramatic whole, treating a single story, as in the *Oresteia,* the only complete trilogy that has survived. His main concern was not dramatic excitement and the portrayal of character but rather the presentation of human action in relation to the overriding purpose of the gods.

His successor was Sophocles, who abandoned for the most part the practice of writing in trilogies, reduced the importance of the chorus, and introduced a third actor. His work too was based on myth, but whereas Aeschylus tried to make more intelligible the working of the divine purpose in its effects on man's life, Sophocles was readier to accept the gods as given and to reveal the values of life as it can be lived within the traditional framework of moral standards. Sophocles' skill in control of dramatic movement and his mastery of speech were devoted to the presentation of the decisive, usually tragic, hours in the lives of men and women at once "heroic" and human, such as Oedipus.

Euripides, last of the three great tragic poets, belonged to a different world. When he came to manhood, traditional beliefs were scrutinized in the light of what was claimed by Sophist philosophers, not always unjustifiably, to be reason; and this was a test to which much of Greek religion was highly vulnerable. The whole structure of society and its values was called into question. This movement of largely destructive criticism was clearly not uncongenial to Euripides. But as a dramatic poet he was bound to draw his material from myths, which, for him, had to a great extent lost their meaning. He adapted them to make room for contemporary problems, which were his real interest. Many of his plays suffer from a certain internal disharmony, yet his sensibilities and his moments of psychological insight bring him far closer than most Greek writers to modern taste. There are studies, wonderfully sympathetic, of wholly unsympathetic actions in the *Medea* and *Hippolytus;* a vivid presentation of the beauty and horror of religious ecstasy in the *Bacchants;* in the *Electra,* a reduction to absurdity of the values of a myth that justifies matricide; in *Helen* and *Iphigenia Among the Taurians,* melodrama with a faint flavour of romance.

Comedy. Like tragedy, comedy arose from a ritual in honour of Dionysus, in this case full of abuse and obscenity connected with averting evil and encouraging fertility. The parabasis, the part of the play in which the chorus broke off the action and commented on topical events and characters, was probably a direct descendant of such revels. The dramatic element may have been derived from the secular Dorian comedy without chorus, said to have arisen at Megara, which was developed at Syracuse by Epicharmus (*c.* 530–*c.* 440). Akin to this kind of comedy seems to have been the mime, a short realistic sketch of scenes from everyday life. These were written rather later by Sophron of Syracuse; only fragments have survived but they were important for their influence on Plato's dialogue form and on Hellenistic mime. At Athens, comedy became an official part of the celebrations of Dionysus in 486 BC. The first great comic poet was Cratinus. About 50 years

later Aristophanes and Eupolis refined somewhat the wild robustness of the older poet. But even so, for boldness of fantasy, for merciless invective, for unabashed indecency, and for freedom of political criticism, there is nothing like the Old Comedy of Aristophanes, whose work alone has survived. Cleon the politician, Socrates the philosopher, Euripides the poet were alike the victims of his masterly unfairness, the first in *Knights*; the second in *Clouds*; and the third in *Women at the Thesmophoria* and *Frogs*; whereas in *Birds* the Athenian democracy itself was held up to a kindlier ridicule. Aristophanes survived the fall of Athens in 404, but the Old Comedy had no place in the revived democracy.

The gradual change from Old to Middle Comedy took place in the early years of the 4th century. Of Middle Comedy, no fully developed specimen has survived. It seems to have been distinguished by the disappearance of the chorus and of outspoken political criticism and by the growth of social satire and of parody; Antiphanes and Alexis were the two most distinguished writers. The complicated plots in some of their plays led to the development of the New Comedy at the end of the century, which is best represented by Menander. One complete play, the *Dyscolus*, and appreciable fragments of others are extant on papyrus. New Comedy was derived in part from Euripidean tragedy; its characteristic plot was a translation into terms of city life of the story of the maiden—wronged by a god—who bears her child in secret, exposes it, and recognizes it years after by means of the trinkets she had put into its cradle.

History. The first great writer of history was Herodotus of Halicarnassus, who was also a geographer and anthropologist. The theme of his history, written in large part for Athenian readers, is the clash between Europe and Asia culminating in the Persian War. The account of the war itself, which occupies roughly the second half of the work, must have been composed by means of laborious inquiry from those whose memories were long enough to recall events that happened when Herodotus was a child or earlier. The whole history, though in places badly put together, is magnificent in its compass and unified by the consciousness of an overriding power keeping the universe and mankind in check.

Thucydides (*c.* 460–*c.* 400) was perhaps the first person to apply a first-class mind to a prolonged examination of the nature of political power and the factors by which policies of states are determined. As a member of the board of generals he acquired inside knowledge of the way policy is shaped. After his failure to save Amphipolis in 424, he spent 20 years in exile, which he used as an opportunity for getting at the truth from both sides. The result was a history of the war narrowly military and political but of the most penetrating quality. Thucydides investigated the effect on individuals and nations both of psychological characteristics and of chance. His findings were interpreted through the many speeches given to his characters.

The histories of Thucydides and Xenophon

Just as Thucydides had linked his work to the point at which Herodotus had stopped, so Xenophon (431–died before 350) began his *Hellenica* where Thucydides' unfinished history breaks off in 411. He carried his history down to 362. His work was superficial by comparison with that of Thucydides, but he wrote with authority of military affairs and appears at his best in the *Anabasis*, an account of his participation in the enterprise of the Greek mercenary army, with which the Persian prince Cyrus tried to expel his brother from the throne, and of the adventurous march of the Greeks, after the murder of their leaders by the Persians, from near Babylon to the Black Sea coast. Xenophon also wrote works in praise of Socrates, of whom his understanding was superficial.

No other historical writing of the 4th century has survived except for a substantial papyrus fragment containing a record of events of the years 396–395. Later, history declined to being merely a province of rhetoric.

Rhetoric and oratory. In few societies has the power of fluent and persuasive speech been more highly valued than it was in Greece, and even in Homer there are speeches that are pieces of finished rhetoric. But it was the rise of democratic forms of government that provided a great incentive to study and instruction in the arts of persuasion, which were equally necessary for political debate in the assembly and for attack and defense in the law courts.

Origin of the study of rhetoric

The formal study of rhetoric seems to have originated in Syracuse *c.* 460 BC with Corax and his pupils Tisias and Gorgias (died *c.* 376); Gorgias was influential also in Athens. Corax is reputed to have been the first to write a handbook on the art of rhetoric, dealing with such topics as arguments from probability and the parts into which speeches should be divided. Most of the Sophists had pretensions as teachers of the art of speaking, especially Protagoras, who postulated that the weaker of two arguments could by skill be made to prevail over the stronger, and Prodicus of Ceos.

Antiphon (*c.* 480–411), the first professional speech writer, was an influential opponent of democracy. Three speeches of his, all dealing with homicide cases, have been preserved, as have three "tetralogies," sets of two pairs of speeches containing the arguments to be used on both sides in imaginary cases of homicide. In them primitive ideas are expressed concerning bloodguilt and the duty of vengeance. Antiphon's style is bare and rather crudely antithetical. Gorgias from Sicily, who visited Athens in 427, introduced an elaborate balance and symmetry emphasized by rhyme and assonance. Thrasymachus of Chalcedon made a more solid contribution to the evolution of a periodic and rhythmical style.

Andocides (*c.* 440–died after 391), an orator who spent much of his life in exile from Athens, wrote three speeches containing vivid narrative; but as an orator he was admittedly amateurish. Lysias (*c.* 455–died after 380) lived at Athens for many years as a resident alien and supported himself by writing speeches when he lost his wealth. His speeches, some of them written for litigants of humble station, show dexterous adaptation to the character of the speaker, though the most interesting of all is his own attack on Eratosthenes, one of the Thirty Tyrants imposed on Athens by the Spartans in 404 BC.

Orators of the 4th century BC

The 12 extant speeches of Isaeus, who was active in the first half of the 4th century BC, throw light on aspects of Athenian law. Isocrates, who was influential in Athens for half a century before his death in 338, perfected a periodic prose style that, through the medium of Latin, was widely accepted as a pattern; and he helped give rhetoric its predominance in the educational system of the ancient world. In his writings, which took the form of speeches but were more like pamphlets, Isocrates shows some insight into the political troubles besetting Greece, with its endless bickering between cities incapable of cooperation.

The greatest of the orators was Demosthenes (384–322), supreme in vehemence and power, though lacking in some of the more delicate shadings of rhetorical skill. His speeches were mainly political, and he is best remembered for his energetic opposition to the rise of Macedonia under its king Philip II, embodied in the three "Philippics." After Demosthenes, oratory faded, together with the political setting to which it owed its preeminence. Two more 4th-century-BC writers need only be mentioned, Hyperides (*c.* 390–322) and Lycurgus (*c.* 390–324).

Philosophical prose. Prose as a medium of philosophy was written as early as the 6th century. Practitioners include Thales, Anaximander, Democritus, and Heracleitus. Philosophical prose was the greatest literary achievement of the 4th century. It was influenced by Socrates (who himself wrote nothing) and his characteristic method of teaching by question and answer, which led naturally to the dialogue. Alexamenus of Teos and Antisthenes, both disciples of Socrates, were the first to use it; but the greatest exponent of Socratic dialogue was the Athenian Plato (428/427–348/347). Shortly after Socrates' death in 399 Plato wrote some dialogues, mostly short; to this group of work belong the *Apology*, *Protagoras*, and *Gorgias*. In the decade after 385 he wrote a series of brilliant works, *Phaedo*, *Phaedrus*, *Symposium*, and the *Republic*. His Socrates is the most carefully drawn character in Greek literature. Subsequent dialogues became more austerely philosophical; Socrates tended increasingly to be a mere spokesman for Plato's thought; and in the last of his works, the *Laws*, he was replaced by a colourless "stranger."

Plato

Plato's style is a thing of matchless beauty, though ancient critics, who were likely to entangle themselves in the rules they had invented, found it too poetical.

Aristotle

Plato's pupil Aristotle (384–322) was admired in antiquity for his style; but his surviving works are all of the "esoteric" sort, intended for use in connection with his philosophical and scientific school, the Lyceum. They are without literary grace, and at times they approximate lecture notes. His works on literary subjects, the *Rhetoric*, and above all, the *Poetics*, had an immense effect on literary theory after the Renaissance. In the ancient world, Aristotelian doctrine was known mainly through the works of his successor Theophrastus (*c.* 372–288/287), now lost except for two books on plants and a famous collection of 30 *Characters*, sketches of human types much imitated by English writers of the 17th century.

With Theophrastus, Attic prose died out for a time. Technical prose in this period was produced in abundance; and rhetoric, a mere literary exercise divorced from political influence, became ornate and flowery in manner.

Late forms of poetry. The creative period of the Hellenistic Age was practically contained within the span of the 3rd century BC. To this period belonged three outstanding poets: Theocritus, Callimachus, and Apollonius of Rhodes. Theocritus (*c.* 310–250), born at Syracuse, is best known as the inventor of bucolic mime, or pastoral poetry, in which he presented scenes from the lives of shepherds and goatherds in Sicily and southern Italy. He also dramatized scenes from middle-class life; and in his second idyll the character Simaetha, who tries by incantations to recover the love of the man who has deserted her, touches the fringe of tragedy. He also used another Hellenistic form, the epyllion, a short scene of heroic narrative poetry in which heroic stature is often reduced by playful realism and delicate psychology. In his hands the hexameter attained a lyric purity and sweetness unrivaled elsewhere. He was the first of the nature poets, succeeded by Moschus and Bion.

The *Aitia* of Callimachus

Callimachus (flourished about 260) was a scholar as well as a poet. His most famous work, of which substantial fragments survive, was the *Aitia*, an elegiac poem describing the origins of various rites and customs. It was heavy with learning but diversified by passages of entertaining narrative. His six hymns show immense poetic expertise but no religious feeling, for the gods of Olympus had long since become obsolete. Callimachus also wrote epigrams, and fragments survive of *iambi* ("iambs"). The form was widely used throughout the 3rd century to denounce the vanities of the world. Sometimes, in a mixture of prose and verse, these pieces had links with satire; and their chief exponents were Bion the Borysthenite, Menippus of Gadara, Cercidas of Megalopolis, and Phoenix of Colophon.

Callimachus avoided epic in favour of the greater intensity possible in shorter works. The last surviving classical Greek epic was written by his successor at Alexandria, Apollonius of Rhodes (born about 295). Apollonius' account of the voyage of the Argonauts is so full of local legend that the coherence of the poem is lost; but the story of Medea's wild passion for Jason, the leader of the Argonauts, is marked by a new sort of romantic awareness that is fully realized in the episode of Dido's passion for Aeneas in Virgil's *Aeneid*.

The desire to combine learning with poetry led to the revival of didactic verse. The *Phaenomena* of Aratus of Soli (*c.* 315–*c.* 245) is a versification of a treatise on the stars by Eudoxus of Cnidus (*c.* 390–*c.* 340). Chance has preserved the poems of Nicander (probably 2nd century) on the unlikely subjects of cures for bites and antidotes to poisons.

The mimes of Herodas (3rd century), short realistic sketches of low life in iambic verse, have affinities with the non-pastoral mimes of Theocritus. They perhaps give a hint as to the character of the literature of popular entertainment, now largely lost. Mime, especially pantomime, was the main entertainment throughout the early Roman Empire.

After the middle of the 3rd century, poetic activity largely died away, though the great period of scholarship at Alexandria and at Pergamum was still to come. The names of a few poets are known: Euphorion (born about 275) of Chalcis and Parthenius (flourished 1st century BC), the teacher of Virgil. Thereafter Greek poetry practically ceased, apart from a sporadic revival in the 4th century AD. An exception exists in the case of epigrammatic poetry in elegiac couplets, surviving mainly in two compilations, the Planudean and Palatine anthologies.

Late forms of prose. Almost all of the great mass of Hellenistic prose—and later prose, historical, scholarly, and scientific—has perished. Among historians Polybius (*c.* 200–*c.* 118 BC), the most outstanding, has survived in a fragmentary condition. Present at Rome when it was succumbing to the first influences of Greek literature, he wrote mainly of events of which he had direct experience, often with great insight; his work covered the period from 264 to 146. Diodorus Siculus' universal history (1st century BC) is important for the sources quoted there. The most considerable of lost historians was Timaeus (*c.* 356–*c.* 260), whose history of the Greeks in the west down to 264 provided Polybius with his starting point. Later historians were Dionysius of Halicarnassus (flourished about 20 BC); Appian of Alexandria (2nd century AD), who wrote on Rome and its conquests; and Arrian (*c.* AD 96–*c.* 180) from Bithynia, who is the most valuable source on Alexander the Great.

The most important works of criticism, of which little has survived, were by Dionysius of Halicarnassus and the obscure Longinus. Longinus' treatise *On the Sublime* (*c.* AD 40) is exceptional in its penetrating analysis of creative literature. The *Bibliotheca* attributed to Apollodorus (*c.* 180 BC) is a handy compendium of mythology.

Scientific work such as the astronomy and geography of Eratosthenes (*c.* 276–*c.* 194) of Alexandria is known mainly from later summaries; but much that was written by the mathematicians, especially Euclid (flourished *c.* 300 BC) and Archimedes (*c.* 287–212), has been preserved.

Much survives of the writings of the physician Galen (AD 129–199). His contemporary Sextus Empiricus is an important source for the history of Greek philosophy. The survey of the Mediterranean by Strabo in the time of Augustus preserved much valuable information; and so, in a more limited field, did the description of Greece by Pausanias (2nd century AD). Greek achievement in astronomy and geography was summed up in the work of Ptolemy of Alexandria in the 2nd century AD.

Greek became the language of the large settlement of Jews at Alexandria, and the Septuagint, the Greek version of the Old Testament, was completed by about the end of the 2nd century BC. Much of the Apocrypha was composed in Greek, and the New Testament was written in popular Greek (Koine). Of the early Christian writers in Greek the most notable were Clement of Alexandria (*c.* AD 150–*c.* 215) and Origen (*c.* AD 185–*c.* 254), together with Clement of Rome and Ignatius of Antioch.

Plutarch's *Lives*

The *Parallel Lives* of famous Greeks and Romans by Plutarch (*c.* AD 46–*c.* 119) of Chaeronea in Boeotia was for centuries one of the formative books for educated Europeans. Great figures from an idealized past are presented for the edification of the lesser men of his own day; and the anecdotes with which the *Lives* abound are of various degrees of credibility. They belong to biography rather than to history, though they are an important source for historians. A number of shorter works on a wide variety of subjects have come down under the Latin title *Moralia* (Greek *Ethica*), which show the intellectual tide of Greece on the ebb.

There was much concern over a question that had been argued ever since the days when Athens had ceased to be a free city: to what extent was Attic prose a norm that writers and especially orators were bound to follow? Many had shunned it in favour of a more ornamental Asiatic style. But at the end of the 1st century AD there was a revival of the Attic dialect. Speeches and essays were written for wide circulation. This revival is known as the Second Sophistic movement, and chief among its writers were Dion Chrysostom (1st century AD), Aelius Aristides (2nd century), and Philostratus (early 3rd century). The only writer of consequence, however, was Lucian (*c.* 120–*c.* 190). His works are mainly slight and satirical; but

his gift of humour, even though repetitive, cannot be denied. *Lives and Opinions of Eminent Philosophers* was a valuable work of the 3rd century by Diogenes Laërtius, a writer otherwise unknown.

Philosophical activity in the early empire was mainly confined to moralizings based on Stoicism, a philosophy advocating a life in harmony with nature and indifference to pleasure and pain. Epictetus (born about AD 55) influenced especially the philosophic Roman emperor Marcus Aurelius (121–180), whose *Meditations* have taken their place beside works of Christian devotion. Many of Plutarch's *Moralia* were Platonic, with vaguely mystical tendencies; but Plotinus (*c.* 205–260/270) was the last major thinker in the classical world, giving new direction to Platonic and Pythagorean mysticism.

Erotic romances

The latest creation of the Greek genius was the novel, or erotic romance. It may have originated as early as the 1st century BC; but its roots reach back to such plays of triumphant love as the lost *Andromeda* of Euripides, to the New Comedy, to Xenophon's daydreams about the education of Cyrus, and to the largely fictitious narratives that were one extreme of what passed for history from the 3rd century BC onward. Of these last, the best known examples are the Alexander romances, a wildly distorted and embroidered version of the exploits of Alexander the Great, which supplied some of the favourite reading of the Middle Ages. Erotic elegy and epigram may have contributed something and so may the lost *Milesian Tales* of Aristides of Miletus (*c.* 100 BC), though these last appear to have depended on a pornographic interest that is almost completely absent from the Greek romances. Only fragments survive of the Ninus romance (dealing with the love of Ninus, legendary founder of Nineveh), which was probably of the 1st century BC; but full-length works survive by Chariton (2nd century AD), Heliodorus (3rd century AD), Xenophon (2nd or 3rd century AD) of Ephesus, and Achilles Tatius (2nd century AD). All deal with true lovers separated by innumerable obstacles of human wickedness and natural catastrophe and then finally united. *Daphnis and Chloe* by Longus (between 2nd and 3rd century AD) stands apart from the others because of its pastoral, rather than quasi-historical, setting. The works of Dictys Cretensis and Dares Phrygius belong to the same period. They claim to give a pre-Homeric account of the Trojan War. The Greek originals are almost wholly lost, but the Latin version was for the Middle Ages the main source for the story of Troy. (D.W.L./Ed.)

Byzantine literature

GENERAL CHARACTERISTICS

Byzantine literature may be broadly defined as the Greek literature of the Middle Ages, whether written in the territory of the Byzantine Empire or outside its borders. By late antiquity many of the classical Greek genres, such as drama and choral lyric poetry, had long been obsolete, and all Greek literature affected to some degree an archaizing language and style, perpetuated by a long-established system of education in which rhetoric was a leading subject. The Greek Church Fathers were the products of this education and shared the literary values of their pagan contemporaries. Consequently the vast and imposing Christian literature of the 3rd to 6th centuries, which established a synthesis of Hellenic and Christian thought, was largely written in a language already far removed from that spoken by all classes in everyday life, and indeed from that of the New Testament. This diglossy—the use of two very different forms of the same language for different purposes—marked Byzantine culture for 1,000 years; but the relations between the high and low forms changed with the centuries. The prestige of the classicizing literary language remained undiminished until the end of the 6th century; only some popular saints' lives and world chronicles escaped its influence. In the ensuing two and a half centuries, when the very existence of the Byzantine Empire was threatened, city life and education declined, and with them the use of classicizing language and style. With the political recovery of the 9th and 10th centuries began a literary revival, in which a conscious attempt was made to recreate the Hellenic-Christian culture of late antiquity. Simple or popular language was despised; many of the early saints' lives were rewritten in inflated and archaizing language and style. By the 12th century the cultural self-assurance of the Byzantines enabled them to develop new literary genres, including romantic fiction, in which adventure and love are the main motifs, and satire, which occasionally made use of imitations of spoken Greek. The period from the Fourth Crusade (1204) to the capture of Constantinople by the Ottoman Turks (1453) saw both a vigorous revival of narrowly imitative, classicizing literature, as the Byzantines sought to assert their cultural superiority over the militarily and economically more powerful West, and at the same time the beginning of a flourishing literature in an approximation to vernacular Greek. But this vernacular literature was limited to poetic romances, popular devotional writing, and the like. All serious writing continued to make use of the prestigious archaizing language of learned tradition.

Relation to classical and Christian literatures

Byzantine literature's two sources, classical and Christian, each provided a series of models and references for the Byzantine writer and reader. Often both were referred to side by side: for example, the emperor Alexius Comnenus defended his seizure of church property to pay his soldiers by referring to the precedents of Pericles and the biblical king David. Much of Byzantine literature was didactic in tone, and often in content too. And much of it was written for a limited group of educated readers, who could be counted upon to understand every classical or biblical allusion and to appreciate every figure of rhetoric. Some Byzantine genres would not be considered of literary interest today, but instead seem to belong to the domain of technical writing. This is true in particular of the voluminous writings of the Church Fathers, such as Athanasius, Gregory of Nazianzus, Basil, John Chrysostom, Cyril of Alexandria, and Maximus the Confessor.

PRINCIPAL FORMS OF WRITING

Nonliturgical poetry. Poetry continued to be written in classical metres and style. But the sense of appropriateness of form to content was lost. An example is the transitional work of Nonnus, a 5th-century Egyptian-born Greek who eventually converted to Christianity. His long poem *Dionysiaca* was composed in Homeric language and metre, but it reads as an extended panegyric on Dionysus rather than as an epic. Nonnus is plausibly credited with a paraphrase, in similar metre and style, of the Gospel According to St. John, thereby fusing classical and Christian traditions. Several short narrative poems in Homeric verse, of mythological content, were composed by contemporaries of Nonnus. Paul the Silentiary in the mid-6th century used the same Homeric form for a long descriptive poem on the Church of the Divine Wisdom (Hagia Sophia) in Constantinople. Many brief occasional poems were written in hexameters or elegiac couplets until the late 6th century. But changes in the phonology of Greek, and perhaps declining educational standards, made these metres difficult to handle. A cleric, George the Pisidian, wrote long narrative poems on the wars of the emperor Heraclius (610–641), as well as a poem on the six days of the creation, in iambic trimeters (12-syllable lines, consisting in principle of three pairs of iambic feet, each of a short syllable followed by a long). His example was followed by Theodosius the Deacon in his epic on the recapture of Crete from the Arabs in the 10th century. This 12-syllable line became the all-purpose metre in the middle and later Byzantine periods and was the vehicle for narrative, epigram, romance, satire, and moral and religious edification. From the 11th century it found a rival in a 15-syllable stressed line, which was used by the monk Symeon the New Theologian in many of his mystical hymns and which became a vehicle for court poetry in the 12th century. It was also used by the metropolitan Constantine Manasses for his world chronicle and by the anonymous redactor of the epic romance of *Digenis Akritas*. It was in this metre, which followed no classical models, that the early vernacular poems were written, such as the romances of *Callimachus and Chrysorrhoe*, *Belthandros and Chrysantza*, the Byzantine *Achilleid* (the

Developments in Byzantine poetry

hero of which has nothing in common with Homer's Achilles but his name), and the *Romance of Belisarius.* These are the most significant works of genuine fiction in Byzantine literature. Many of these poems were adaptations or imitations of medieval Western models: examples are *Phlorios and Platziaphlora* (the Old French *Floire et Blancheflor*), *Imberios and Margarona,* and *Apollonius of Tyre,* each a romantic narrative. The epic genre is represented by a long unpublished poem on the Trojan War, adapted from the *Roman de Troie* of the 12th-century French poet Benoît de Sainte-Maure. This openness to the Latin West was new. But even when they were based on Western models, Byzantine poems differed in tone and expression from their exemplars. Most of this vernacular poetry cannot be dated more precisely than to the 13th or 14th century.

Much Byzantine poetry is rather unimaginative, long-winded, and tedious. But some poets show a genuine vein of inspiration, for instance, John Geometres (10th century) or John Mauropous (11th century), or remarkable technical brilliance, such as Theodore Prodromus (12th century), or Manuel Philes (14th century). The ability to write passable verse was widespread in literate Byzantine society, and poetry—or versification—was greatly appreciated.

Liturgical poetry. From the earliest times song—and short rhythmic stanzas (*troparia*) in particular—had formed part of the liturgy of the church. Poems in classical metre and style were composed by Christian writers from Clement of Alexandria and Gregory of Nazianzus to Sophronius of Jerusalem. But the pagan associations of the genre, as well as the difficulties of the metre, made them unacceptable for general liturgical use. In the 6th century elaborate rhythmical poems (*kontakia*) replaced the simpler *troparia.* They owed much to Syriac liturgical poetry. In form the *kontakion* was a series of up to 22 rhythmical stanzas, all constructed on the same accentual pattern and ending with the same short refrain. In content it was a narrative homily on an event of biblical history or an episode in the life of a saint. There was often a marked dramatic element. Rich in imagery, complex in structure, and infinitely variable in rhythm, the new liturgical poetry can be compared with the choral lyric of ancient Greece. The greatest composer of *kontakia* was Romanos Melodos (Romanos the Melode; early 6th century), a Syrian probably of Jewish origin. In the late 7th century the *kontakion* was replaced by a longer liturgical poem, the *kanōn,* consisting of eight or nine odes, each of many stanzas and each having a different rhythmic and melodic form. The *kanōn* was a hymn of praise rather than a homily. Its great length encouraged repetition and inflation, and a more ornamental style of singing enhanced the importance of the music at the expense of the words. The most noteworthy composers of *kanones* were Andrew of Crete, John of Damascus, Theodore Studites, Joseph the Hymnographer, and John Mauropous. No new hymns were added to the liturgy after the 11th century, but *kanones* continued to be composed as a literary exercise. The original music of *kontakia* and *kanones* alike is lost.

The *kontakion*

The *kanōn*

Historical works. Conscious as they were of their classical and biblical past, the Byzantines wrote much history. Until the early 7th century a series of historians recounted the events of their own time in classicizing style, with fictitious speeches and set descriptive pieces, in a genre that owed much to the classical Greek historians Thucydides and Polybius. Procopius, Agathias, Peter the Patrician, Menander Protector, and Theophylactus Simocattes each took up where a predecessor left off. Thereafter this vein virtually ran dry for 250 years. The revival of cultural confidence and political power in the late 9th century saw a revival of classicizing history, with an interest in human character—Plutarch was often the model—and the causes of events. Joseph Genesius in the 10th century and the group of historical writers known collectively as the Continuators of Theophanes recorded, not without partiality, the origin and early days of the Macedonian dynasty. From then until the later 14th century there was never a generation without its historian. The most noteworthy historians were Symeon the Logothete and Leo the Deacon in the 10th century; Michael Psellus, Michael Attaleiates, and John Scylitzes in the 11th century; Anna Comnena, John Cinnamus, and Nicetas Choniates in the 12th century; George Acropolites and George Pachymeres in the 13th century; and Nicephorus Gregoras and the emperor John Cantacuzenus in the 14th century. The last days of the Byzantine Empire were recounted from very different points of view by George Sphrantzes, the writer known simply as Ducas (who was a member of the former Byzantine imperial house of that name), Laonicus Chalcocondyles, and Michael Critobulus in the second half of the 15th century.

Another kind of interest in the past was satisfied by world chronicles, beginning with the creation or some early biblical event. Often naively theological in their explanation of causes, black-and-white in their depiction of character, and popular in language, they helped the ordinary Byzantine to locate himself in a scheme of world history that was also a history of salvation. The *Chronographia* of John Malalas in the 6th century and the Paschal Chronicle (*Chronicon Paschale*) in the 7th century were succeeded by those of Patriarch Nicephorus at the end of the 8th century, Theophanes the Confessor in the early 9th century, and George the Monk in the late 9th century. Such chronicles continued to be written in later centuries, sometimes with critical and literary pretensions, as in the case of John Zonaras, or in vaguely romanticized form in verse, as in the case of Constantine Manasses.

World chronicles

The importance that Byzantine rulers attached to history is attested by the vast historical encyclopaedia compiled on the orders of Constantine VII (913–959) in 53 volumes, of which only meagre fragments remain.

Rhetoric. Though there was no opportunity for political or forensic oratory in the Byzantine world, the taste for rhetoric and the appreciation of well-structured language, choice figures of speech and thought, and skillful delivery remained undiminished in Byzantine society. From the 10th century onward survives a vast body of encomiums, funeral orations, memorial speeches, inaugural lectures, addresses of welcome, celebrations of victory, and miscellaneous panegyrics. This outpouring of polished rhetoric played an important role in the formation and control of public opinion in the limited circles where opinions mattered and occasionally served as a vehicle of genuine political controversy. To this same domain belong the myriad Byzantine letters, often collected and edited by their author or a friend. These letters were not intended to be either private or informative—real information was conveyed orally by the bearer—but they were important in maintaining networks of contact among the elite as well as in providing refined aesthetic pleasure. (R.B.)

Modern Greek literature (after 1453)

POST-BYZANTINE PERIOD

After the Turkish capture of Constantinople in 1453, Greek literary activity continued almost exclusively in those areas of the Greek world under Venetian rule. Thus Cyprus, until its capture by the Turks in 1571, produced a body of literature in the local dialect, including the 15th-century prose chronicle *Recital Concerning the Sweet Land of Cyprus* by Leóntios Machairás and a collection of translations and imitations in elaborate verse forms of Italian poems by Petrarch and others. Crete, which remained in Venetian hands until 1669, became the centre of the greatest flowering of Greek literature between the fall of Constantinople and the foundation of the modern Greek state. There a number of authors developed the Cretan dialect into a rich and subtle medium of expression. In it were written a number of tragedies and comedies, a single pastoral tragicomedy, and a single, anonymous religious drama, *The Sacrifice of Abraham,* mostly based on Italian models. The leading playwright was Geórgios Chortátsis. In the first half of the 17th century Vitsḗntsos Kornáros composed his romance, *Erotókritos.* These Cretan authors composed their works almost entirely in the 15-syllable iambic verse of the Greek folk song, whose modes of expression influenced them deeply.

In the Ottoman-ruled areas of Greece the folk song, which concisely and unsentimentally conveyed the aspira-

tions of the Greek people of the time, became practically the sole form of literary expression.

Toward the end of the 18th century, however, a number of intellectuals emerged who, under the influence of European ideas, set about raising the level of Greek education and culture and laying the foundations of an independence movement. The participants in this "Greek Enlightenment" also brought to the fore the language problem, each promoting a different form of the Greek language for use in education. The leading Greek intellectual of the early 19th century was the classical scholar Adamántios Korais, who in voluminous writings on Greek language and education, argued for a form of Modern Greek "corrected" according to the ancient rules.

The Greek language problem

INDEPENDENCE AND AFTER

Old Athenian School. The Greek state established as a result of the Greek War of Independence (1821–29) consisted only of a small section of the present-day Greek mainland and a few islands. Athens, which became the capital of Greece in 1834, soon came to be the chief cultural centre, gathering together writers from various areas, particularly Constantinople. The Soútsos brothers, Aléxandros and Panayótis, introduced the novel into Greece, but they are best known for their Romantic poetry, which as time went by moved gradually away from the Demotic ("popular"), or commonly spoken, language toward the Katharevusa ("purist") form institutionalized by Korais. The work of these writers, which relied greatly on French models, looks back to the War of Independence and the glorious ancient past. Their melancholy sentimentality was not shared by Aléxandros Rízos Rangavís, a verbose but versatile and not inconsiderable craftsman of Katharevusa in lyric and narrative poetry, drama, and the novel. By the 1860s and '70s, however, Athenian poetry was generally of poor quality and was dominated by a sense of despair and longing for death. Prose throughout the period was monopolized by the historical novel. Emmanuel Roídis' novel called *I Pápissa Ioánna* (1866; *Pope Joan*) is a hilarious satire on medieval and modern religious practices as well as a pastiche of the historical novel. Pávlos Kalligás, in *Thános Vlékas* (1855), treated contemporary problems such as brigandage. In *Loukís Láras* (1879; Eng. trans., *Loukis Laras*) Dimítrios Vikélas presented a less heroic view of the War of Independence.

Heptanesian School. Meanwhile more interesting developments had been taking place in the Ionian Islands (Heptanesos). During the 1820s two poets from the island of Zacynthus made their name with patriotic poems celebrating the War of Independence. One of these, Andréas Kálvos, who composed his odes in neoclassical form and archaic language, never wrote poetry afterward, while the other, Dhionísios Solomós, went on to become one of the greatest of modern Greek poets. Dealing with the themes of liberty, love, and death, Solomós embodied a profoundly Romantic sensibility in extraordinary fragments of lyrical intensity, which gave a new prestige to the Demotic language. Solomós' followers continued to cultivate the Demotic, particularly Antónios Mátesis, whose historical social drama, *O vasilikós* (1859; "The Basil Plant"), was the first prose work of any length to be written in the Demotic. Aristotélis Valaorítis continued the Heptanesian tradition with long patriotic poems inspired by the Greek national struggles.

Demoticism and folklorism, 1880–1922. From the 1880s onward the New Athenian School, inspired by the revived interest in folklore as a survival of ancient Greek culture, began to react against the sterile bombast of the Katharevusa versifiers, producing instead a more intimate poetry based on the language, customs, and beliefs of the Greek peasantry, and in particular on Greek folk songs.

The leading ideologist of this "demoticist" movement, which aimed to promote traditional popular culture at the expense of the pseudo-archaic pedantry fashionable in Athens, was Yánnis Psicháris (Jean Psichari), whose book *My Journey* (1888) was partly a fictionalized account of a journey around the Greek world and partly a belligerent manifesto arguing that the Demotic language should be officially adopted as a matter of national urgency. The demoticist movement inspired poets to enrich the Greek popular tradition with influences from abroad. Chief among these was Kostís Palamás, who dominated the literary scene for several decades with a large output of essays and articles and whose best poetry appeared between 1900 and 1910. In his lyric and epic poems he attempted to synthesize ancient Greek history and mythology with the Byzantine Christian tradition and modern Greek folklore in order to demonstrate the essential unity of Greek culture. Angelos Sikelianós continued this enterprise in effusive and powerful lyric poetry of a profoundly mystical nature.

In prose, the folklore cult fostered development of the short story, written initially in Katharevusa, with Demotic gradually taking over in the 1890s. These stories, and the novels that accompanied them, depicted scenes of traditional rural life, sometimes idealized and sometimes viewed critically by their authors. The pioneer of the Greek short story, Geórgios Vizyenós, combined autobiography with an effective use of psychological analysis and suspense. The most famous and prolific short-story writer, Aléxandros Papadiamándis, produced a wealth of evocations of his native island of Skiáthos imbued with a profound sense of Christian tradition and a compassion for country folk; his novel *I fónissa* (1903; *The Murderess*) is a fine study in psychological abnormality. The novel *O zitiános* (1896; *The Beggar*), by Andréas Karkavítsas, satirically depicts the economic and cultural deprivation of the rural population. From about 1910 this critical attitude is further reflected in the prose writing of Konstantínos Chatzópoulos and Konstantínos Theotókis. Meanwhile Grigórios Xenópoulos wrote novels with an urban setting and devoted considerable effort to drama, a medium that received a substantial boost from the demoticist movement.

The short story

One major figure defies categorization for it was outside Greece, in Alexandria, that Constantine Cavafy lived and wrote. His finely wrought, epigrammatic poems, with their tragically ironic views of Hellenistic and Byzantine history, contain daring, sensuous glimpses of homosexual love.

Literature from 1922. The Asia Minor Disaster of 1922, in which Greece's expansionist designs on the Ottoman Empire were finally thwarted, brought about a radical change in the orientation of Greek literature. Before committing suicide, Kóstas Kariotákis wrote some bitterly sarcastic poetry conveying the gap between the old ideals and the new reality.

The reaction against the defeatism of 1922 came with the Generation of 1930, a group of writers who began publishing around that date. They reinvigorated Greek literature by discarding the old verse forms in poetry and by producing ambitious novels that were intended to embody the spirit of the times. Both poets and novelists sought to combine European influences with the best of what was Greek. The restrained poetry of George Seféris skillfully married references to ancient mythology with pensive meditation on man's modern situation, while his finely written essays recast the Greek tradition according to his own priorities. Odysseus Elýtis celebrated the Aegean scenery as an ideal world of sensual enjoyment and moral purity. Each of these poets won the Nobel Prize for Literature, Seféris in 1963 and Elýtis in 1979. Yánnis Ritsos adopted various new modes of writing in his celebration of the Greek partisans in World War II, in long dramatic monologues spoken by characters from Greek mythology, and in laconic poems depicting everyday, but often ironically presented, scenes.

The Generation of 1930 produced some remarkable novels, among them Strátis Myrivílis' *I zoí en tafo* (1930; *Life in the Tomb*), a journal of life in the trenches in World War I; *Argo* (2 vol., 1933 and 1936) by Yórgas Theotokás, about a group of students attempting to find their way through life in the turbulent 1920s; and *Eroica* (1937) by Kosmás Politis, about the first encounter of a group of well-to-do schoolboys with love and death.

After World War II prose writing was dominated by novels reflecting the experiences of the Greeks during eight years of war (1941–49). Iánnis Beráits recounted his experiences of 1941 in an unemotional manner in *To Platy*

World War II and afterward

Potamí (1946; "The Broad River"). In a trilogy of novels entitled *Akyvérnites polities* (1960–65; *Drifting Cities*), Stratís Tsirkas masterfully recreated the atmosphere of the Middle East in World War II. In the short story, Dimítris Chatzís painted ironic portraits of real and fictional characters in his native Ioánnina in the period before and during World War II, exposing their self-interested machinations.

Kazantzákis' work

Nevertheless, the most famous novelist of the period, the Cretan Níkos Kazantzákis, was a survivor from an earlier generation. In a series of novels, beginning with *Víos ke politía tou Aléksi Zorbá* (1946; *Zorba the Greek*) and continuing with his masterpiece *O Christos xanastavronete* (1954; *Christ Recrucified*), he embodied a synthesis of ideas from various philosophies and religions in larger-than-life characters who wrestle with great problems, such as the existence of God and the purpose of human life. Kazantzákis had earlier published his 33,333-line *Odísia* (1938; *Odyssey*), an epic poem taking up the story of Odysseus where Homer had left off. Pandelís Prevelákis published a number of philosophical novels set in his native Crete, the most successful being *O ílios tou thanátou* (1959; *The Sun of Death*), which shows a boy learning to come to terms with death.

During the 1960s Greek prose writers attempted to explore the historical factors underlying the contemporary social and political situation. In the novel *To tríto stefáni* (1962; *The Third Wedding*) by Kóstas Tachtsís, the female narrator tells the story of her life with venomous verve, unwittingly exposing the oppressive nature of the Greek family. Yórgos Ioánnou's part-fictional, part-autobiographical short prose pieces present a vivid picture of life in Thessaloníki (Salonika) and Athens from the 1930s to the 1980s.

No individual poets of the postwar generations tower above the rest; but Tákis Sinópoulos, Míltos Sachtoúris, and Manólis Anagnostákis, all marked by their wartime experiences of the 1940s, are among those with the greatest reputations. (P.A.M.)

BIBLIOGRAPHY

Ancient Greek literature: Among many surveys of Greek literature, the best are P.E. EASTERLING and B.M.W. KNOX, *The Cambridge History of Classical Literature,* vol. 1, *Greek Literature* (1985); JACQUELINE DE ROMILLY, *A Short History of Greek Literature* (1985; originally published in French, 1980), with an excellent bibliography; K.J. DOVER (ed.), *Ancient Greek Literature* (1980); and ALBIN LESKY, *A History of Greek Literature* (1966; originally published in German, 2nd ed., 1963). Still valuable are GILBERT MURRAY, *A History of Ancient Greek Literature* (1897, reprinted 1966); and MOSES HADAS, *A History of Greek Literature* (1950, reissued 1962). N.G.L. HAMMOND and H.H. SCULLARD (eds.), *The Oxford Classical Dictionary,* 2nd ed. (1970, reprinted 1977); and PAUL HARVEY (comp.), *The Oxford Companion to Classical Literature* (1937, reprinted with corrections 1980), are excellent reference resources. Topical studies include ROBERT GRAVES, *The Greek Myths,* 2 vol. (1955, reprinted with amendments, 2 vol. in 1, 1957); J.W.H. ATKINS, *Literary Criticism in Antiquity: A Sketch of Its Development,* 2 vol. (1934, reissued 1961); C.M. BOWRA, *Greek Lyric Poetry from Alcman to Simonides* (1961, reprinted 1967); MARGARETE BIEBER, *The History of the Greek and Roman Theater,* 2nd rev. ed. (1961); BRIAN VICKERS, *Comparative Tragedy,* vol. 1, *Towards Greek Tragedy: Drama, Myth, Society* (1973, reprinted 1979); R.R. BOLGAR, *The Classical Heritage and Its Beneficiaries* (1954, reprinted 1977); GILBERT HIGHET, *The Classical Tradition: Greek and Roman Influences on Western Literature* (1949, reissued 1985); and M.J. ANDERSON (ed.), *Classical Drama and Its Influences* (1965). Trends in scholarship are discussed in MAURICE PLATNAUER, *Fifty Years (and Twelve) of Classical Scholarship,* 2nd ed. (1968).

(D.W.L./Ed.)

Byzantine literature: The standard reference works are HANS-GEORG BECK, *Kirche und theologische Literatur im byzantinischen Reich,* 2nd ed. (1977), and *Geschichte der byzantinischen Volksliteratur* (1971); and HERBERT HUNGER, *Die hochsprachliche profane Literatur der Byzantiner,* 2 vol. (1978). The standard work on Byzantine liturgical poetry is EGON WELLESZ, *A History of Byzantine Music and Hymnography,* 2nd ed. rev. and enlarged (1961, reprinted 1971). GEORGE A. KENNEDY, *Greek Rhetoric Under Christian Emperors* (1983), is concerned with social aspects, particularly with the rhetorical method in education. A collection of studies on the vernacular poetry of the period is E.M. JEFFREYS and M.J. JEFFREYS, *Popular Literature in Late Byzantium* (1983). ALEXANDER KAZHDAN, *Studies on Byzantine Literature of the Eleventh and Twelfth Centuries* (1984), brings methods of Western medievalists to the study of Byzantine literature. Influences are traced in KENNETH M. SETTON, *The Byzantine Background to the Italian Renaissance* (1956), a reliable survey. Also useful is *Ihor Ševčenko, Ideology, Letters and Culture in the Byzantine World* (1982).

(R.B.)

Modern Greek literature: C.TH. DIMARAS (K.TH. DEEMARAS), *A History of Modern Greek Literature* (1972; originally published in Greek, 2 vol., 1948–49), is a comprehensive study that is especially useful on the history of ideas. LINOS POLITIS, *A History of Modern Greek Literature* (1973), is the best general survey. Other useful works include CONSTANTINE A. TRYPANIS, *Greek Poetry: From Homer to Seferis* (1981), covering the entire Greek poetic tradition; RODERICK BEATON, *Folk Poetry of Modern Greece* (1980); PHILIP SHERRARD, *The Marble Threshing Floor: Studies in Modern Greek Poetry* (1956, reprinted 1981); EDMUND KEELEY and PETER BIEN (eds.), *Modern Greek Writers: Solomos, Calvos, Matesis, Palamas, Cavafy, Kazantzakis, Seferis, Elytis* (1972); EDMUND KEELEY, *Modern Greek Poetry: Voice and Myth* (1983); and PETER BIEN, *Three Generations of Greek Writers: Introductions to Cavafy, Kazantzakis, Ritsos* (1983). See also MARY P. GIANOS (comp.), *Introduction to Modern Greek Literature: An Anthology of Fiction, Drama, and Poetry* (1969).

(P.A.M.)

Biological Growth and Development

Although the essence of growth is an increase in the size or the amount of an entity, it is useful to distinguish between growth of living things and that of inanimate objects. The growth of inanimate objects, such as the process that occurs during the formation of crystals, is limited in an important way. The crystals can increase in size but are unable to reproduce themselves; living things, on the other hand, not only increase in size but also reproduce themselves. Growth of living things always occurs by an increase in the number or size of the basic units of organisms, cells, and always eventually includes reproduction. To appreciate fully the way in which living things grow and the ways this growth is regulated requires a consideration of the way in which cells increase in size and number.

The word growth is often used synonymously with development, and, indeed, increase in size is a striking feature of development. But development, in this context, may also include changes in the types of cell specialization (differentiation) and extensive movements of cells. Movements of cells and tissues during embryonic development are partially responsible for the form of the body and various organs.

This article is concerned with the processes of growth and development of living organisms from the moment that fertilization takes place through the processes of maturation and senescence. (Death in the biological sense and its other aspects are treated in detail in the article DEATH.)

This article is divided into the following sections and subsections:

Growth 362
- The process of growth 362
  - Types of growth
  - Normal and abnormal growth
- Factors that regulate growth 363
  - Environmental factors
  - Internal factors
- The dynamics of growth 364
  - Measurement of growth
  - The study of growth
- Aberrations of growth: biological malformation 365
  - Plant malformations
  - Animal malformations
- Biological regeneration 367
  - Modes of regeneration
  - The regeneration process
  - The range of regenerative capability

General features of biological development 371
- The scope of development 372
  - Types of development
  - General systems of development
  - Constituent processes of development
- Control and integration of development 376
  - Phenomenological aspects
  - Analytical aspects
- Development and evolution 377
  - Effect on life histories
  - Genetic assimilation

Plant development 379
- General features 379
  - Life cycles
  - Body plans
  - Preparatory events
- Early development: from zygote to seedling 381
  - Embryo formation
  - Germination and early growth
- Later development: the sporophyte plant body 384
  - Continuation of organ formation
  - The shoot system and its derivatives
  - The root system and its derivatives
- Correlations in plant development 388
  - Coordination of shoot and root development
  - Determination of mature form
  - Seasonal adaptations

Animal development 390
- General features 390
  - Reproduction and development
  - Preparatory events
- Early development 391
  - Embryo formation
  - Embryonic adaptations
- Organ formation 396
  - Primary organ rudiments
  - Organogenesis and histogenesis
  - Ectodermal derivatives 398
    - The nervous system
    - The epidermis and its outgrowths
  - Mesodermal derivatives 400
    - The body muscles and axial skeleton
    - The appendages: tail and limbs
    - Excretory organs
    - Circulatory organs
    - Reproductive organs
  - Endodermal derivatives 404
    - The alimentary canal
    - The pharynx and its outgrowths
    - The liver, pancreas, and lungs
  - Postembryonic development 404
    - The larval phase and metamorphosis
    - Direct development
  - Maturity and death 407

Human growth and development 407
- Human embryology: early stages 407
  - From fertilization to placentation
  - Formation of the three primary germ layers
  - Growth and differentiation
  - Embryonic acquisition of external form
- Abnormal development 413
  - Multiple births
  - Fetal deviations
  - Teratology
- Development of organs 414
  - Ectodermal derivatives
  - Mesodermal derivatives
  - Endodermal derivatives
- Postnatal development 418
  - Types and rates of human growth
  - Types of growth data
- Development at puberty 421
  - Alterations in growth rate
  - Normal variations
- Hormones and growth 424

Aging and senescence 424
- Life-span 424
  - Measurement of life-span
  - Plants
  - Animals
  - Human life-span
- Aging: general considerations 429
  - Biological theories of aging
  - Natural history of aging
  - Senescence in plants
  - Senescence in mammals
  - Aging at the molecular and cellular levels
  - Internal and external causes of aging
- Human aging 435
  - Effect of aging of the body systems
  - Psychological aspects of aging

Bibliography 439

GROWTH

The process of growth

The increases in cell size and number that take place during the life history of an organism are seldom random; rather, they occur according to a plan that eventually determines the size and shape of the individual. Growth may be restricted to special regions of the organism, such as the layers of cells that divide and increase in size near the tip of the plant shoot. Or the cells engaged in growth may be widely distributed throughout the body of the organism, as in the human embryo. In the latter case, the rates of cell division and of the increase in cell size differ in different parts. That the pattern of growth is predetermined and regular in plants and animals can be seen in the forms of adults. In some organisms, however, notably the slime molds (see PROTOPHYTES: *Slime molds*), no regular pattern of growth occurs, and a formless cytoplasmic mass is the result.

The rate of growth of various components of an organism may have important consequences in its ability to adapt to the environment and hence may play a role in evolution. For instance, an increase in the rate of growth of fleshy parts of the fish fin would provide an opportunity for the fish to adapt more easily to terrestrial locomotory life than could a fish without this modified fin. Without disproportionate growth of the fin—ultimately resulting from random changes in the genetic material (mutations)—the evolution of limbs through natural selection might have been impossible.

TYPES OF GROWTH

In cells. The increase in size and changes in shape of a developing organism depend on the increase in the number and size of cells that make up the individual. Increase in cell number occurs by a precise cellular reproductive mechanism called mitosis. During mitosis the chromosomes bearing the genetic material are reproduced in the nucleus, and then the doubled chromosomes are precisely distributed to the two daughter cells, one of each chromosomal type going to each daughter cell. Each end of the dividing cell receives a complete set of chromosomes before the ends separate. In animal cells this is a pinching off (cytokinesis) of the cell membrane; in plant cells a new cellulose wall forms between the new cells.

During the period of cell life preceding the actual distribution of chromosomes, the mother cell often grows to twice its original size. Hence, a cycle consisting of cell growth and cell division is established. Cell growth—an increase in cytoplasmic mass, chromosome number, and cell surface—is followed by cell division, in which the cytoplasmic mass and chromosomes are distributed to the daughter cells. An increase in cytoplasmic mass does not always occur during cell-division cycles, however. During the early development of an embryo, for example, the original egg cell, usually a very large cell, undergoes repeated series of cell divisions without any intervening growth periods; as a result, the original egg cell divides into thousands of small cells. Only after the embryo can obtain food from its environment does the usual pattern of growth and mitosis occur.

In plants. The fact that most plant cells undergo extensive size increase unaccompanied by cell division is an important distinction between growth in plants and in animals. Daughter cells arising from cell division behind the tip of the plant root or shoot may undergo great increases in volume. This is accomplished through uptake of water by the cells; the water is stored in a central cavity called a vacuole. The intake of water produces a pressure that, in combination with other factors, pushes on the cellulose walls of the plant cells and is responsible for the increase in length and girth of the cells and of the plant. In plants, much of the size increase occurs after cell division and results primarily from an increase in water content of the cells without much increase in dry weight.

The role of water in plant-cell growth

The very young developing plant embryo has many cells distributed throughout its mass that undergo the cycle of growth and cell division. As soon as the positions of the root tip, shoot tip, and embryonic leaves become established, however, the potential for cell division becomes restricted to cells in certain regions called meristems. One meristematic centre lies just below the surface of the growing root; all increases in the number of cells of the primary root occur at this point. Some of the daughter cells remain at the elongating tip and continue to divide. Other daughter cells, which are left behind in the root, undergo the increase in length that enables the new root to push deeper into the soil. The same general plan is evident in the growing shoot of higher plants, in which a restricted meristematic region at the tip is responsible for the formation of the cells of the leaves and stem; cell elongation occurs behind this meristematic centre. The young seedling secondarily develops cells associated with the vascular strands of phloem and xylem—tissues that carry water to the leaves from the soil and sugar from the leaves to the rest of the plant. These cells can divide again, providing new cell material for development of a woody covering and for more elaborate vascular strands. Hence, the growth of higher plants—*i.e.*, those aspects involving both the pattern of stems, leaves, and roots and the increase in bulk—results primarily from cell division at the meristem followed by a secondary increase in size because of water uptake. These activities occur throughout the period of plant growth.

In animals. The growth of animals is more restricted in time than is that of plants, but cell division is more generally distributed throughout the body of the organism. Although the rate of cell division differs in different regions, the capacity for cell division is widely distributed in the developing embryo. Increase in size is rapid during the embryonic period, continues at a reduced rate in juveniles, and thereafter is absent. Cell division and size increase continue, however, even after increase in total body size no longer occurs. Because these events are balanced by cell death, post-juvenile increase in cell number is primarily a replacement phenomenon. Height increase in mammals is limited by cessation of cell division and bone deposition in the long bones. The long juvenile period of growth in humans is unusual, most higher animals attaining mature size soon after the end of embryonic development. Some organ systems undergo little cell division and growth after birth; for instance, all of the germ cells (precursors of egg cells) of the female are formed by the time of birth. Similarly, all of the nerve cells of the brain are formed by the end of the embryonic period. Further increase in the size of the nervous system occurs by outgrowth of nerve fibres and deposition of a fatty insulation material along them. Although the greatest increase in size of nerve cells occurs, as in plant cells, after the cessation of cell division, the nerve fibre outgrowth in animals represents a true increase in the amount of cytoplasm and cell surface and not just an uptake of water.

Some organs retain the potential for growth and cell division throughout the life-span of the animal. The liver, for example, continues to form new cells to replace senescent and dying ones. Although cell division and growth occur throughout the liver, other organs have a special population of cells, called stem cells, that retain the capacity for cell division. The cells that produce the circulating red cells of mammalian blood are found only in the marrow of the long bones. They form a permanent population of dividing cells, replacing the red cells that continuously die and disappear from the circulation.

The rates of both growth and cell division can vary widely in different body parts. This differential increase in size is a prime factor in defining the shape of an organism.

NORMAL AND ABNORMAL GROWTH

Tumours. When growth is not properly regulated, anomalies and tumours may result. If the increase in the number of liver cells is abnormal, for example, tumours

Features associated with tumour growth

of the liver, or hepatomas, may result. In fact, one feature of malignant tumours, or cancers, is the absence of the usual growth patterns and rates. The cells of malignant tumours, in addition to having abnormal growth rates, have altered adhesive properties, which enable them to detach easily from the tumour; in this way the cells may spread to other parts of the body (metastasize) and grow in unusual locations. It is the growth of tumours in places other than the organ of origin that usually causes the death of an organism. Tumours may vary widely in their growth rates. They may grow very rapidly or so slowly that the rate approaches that of normal cell division in adult tissues. Tumours are not only characterized by an increase in the rate of cell division but also by abnormal patterns of growth. The new cells formed in the tumour are not organized and incorporated into the structure of the organ and may form large nodules. These abnormal growths may present no medical problems (*e.g.*, moles) or may cause disastrous effects, as is the case of the pressure on the brain caused by a tumorous mass of the meningeal covering of the brain. (For additional information about abnormal growth, see below *Aberrations of growth: biological information.*)

Regeneration. Not all abnormal growths are tumours. If a tree is partially burned, cells below the bark produce a new covering for the exposed vascular strands. Growth may not be normal, and an obvious scar or growth of the new bark is apparent. Similarly, if the skin of a mammal is severely injured, the repair, although abnormal and imperfect, causes the organism no physiological difficulty. Many organisms possess the ability to regrow, or regenerate, with varying degrees of perfection, parts of the body that are lost or injured. Salamanders possess remarkable powers of regeneration, being able to form new eyes or a new limb if the original is lost. Lizards can regenerate a new tail; even humans can regenerate parts of the liver. The reasons for the differences in regenerative powers in different animals remain a fascinating mystery of great practical importance. When regeneration does occur, some specialized cells usually lose their specialized characteristics and enter a period of an increased rate of cell division; subsequently, the new cells respecialize into the tissues of the original body part. Plants whose tops are lost as in pruning can also sometimes form new meristematic centres from dormant tissues and produce new shoots. (For additional information see below *Biological regeneration.*)

Compensatory growth. Many organs of animals occur in pairs, and if one is lost the remaining member increases in size, as if responding to the demands of increased use. If one of the two kidneys of a human is removed, for example, the other increases in size. This is called a compensatory reaction and may occur either by some increase in cell size (hypertrophy), by an increase in the rate of cell division (hyperplasia), or both. Although an increase in cell number is primarily responsible for the compensatory reaction of the kidney, the number of individual filtration units (glomeruli) does not increase. Hence, cell division increases the size of glomeruli but not the total number. Some of the most striking examples of increases in cell size in animals take place during stimulation of endocrine organs, which secrete regulatory substances called hormones; when the thyroid gland is stimulated, for example, the individual cells of the gland may increase dramatically in size.

Compensatory growth in the kidney

Factors that regulate growth

ENVIRONMENTAL FACTORS

Temperature. The environment in which an organism lives plays an important role in modifying the rate and extent of growth. Environmental factors may be either physical (*e.g.*, temperature, radiant energy, and atmospheric pressure) or chemical. Organisms and the cells of which they are composed are extremely sensitive to temperature changes; as the temperature decreases, the biochemical reactions necessary for life occur more slowly. A lowering of the temperature by 10° C (18° F) slows metabolism at least twofold and often more.

The width of trees increases partly by cell division and enlargement of secondary meristematic tissue below the bark. During the cold of winter, cell division and enlargement may cease completely; but during the spring renewed growth occurs. This intermittent growth is influenced by temperature, light, and water. The amount of growth may decrease considerably if the spring is cold, if day length is changed by obstructions blocking the sunlight, or if a drought occurs. In fact, the width of the growth rings visible on the surface of the cut tree trunk provides a partial history of climatic conditions, the spacing of the growth rings of different size having been correlated with known periods of drought and cold to provide reliable archaeological dating of various structures, as in the timbers used in Indian pueblos in the southwestern United States.

Temperature also affects both warm- and cold-blooded animals. Many warm-blooded (*e.g.*, bears) and cold-blooded (*e.g.*, frogs) vertebrates cease growing during the cold winter and simply enter an inactive or dormant state, which is characterized by a very low rate of metabolism. In animals that do not become dormant, increased demands for food consumption occur during cold periods to provide energy to maintain body temperature; this utilization of food energy may limit the energy available for size increase if food is in short supply.

Dormancy

Pressure. Because atmospheric pressure is relatively constant except in the mountains, it probably is of little importance in growth regulation. Increases in pressure in the ocean's depths may be significant, however, since it is known that increases in hydrostatic pressure interfere with cell division. Tissues of deep-sea fishes must have become adapted to such pressure effects, which have been little studied thus far. Movements of the terrestrial atmosphere—winds—may affect growth patterns in trees and shrubs, as is evident in the exotic shapes of certain conifers that grow along coastlines exposed to strong prevailing winds.

Light. Of all the physical factors, light plays the best understood and most dramatic role. Many of the effects of light on plant growth are obvious and direct. Light energy is the driving force for photosynthesis, the series of chemical reactions in green plants in which carbon dioxide and water form carbohydrates and upon which all life ultimately depends. Insufficient light causes death or retardation of growth in green plants. But light also has indirect effects of great importance. Green plants possess small amounts of a pigment called phytochrome that can exist in two forms. One form absorbs red light (660 millimicrons, or mμ; 1 mμ = 3.937×10^{-8} inch). When plants containing this pigment absorb red light, the pigment is converted to another form, which absorbs far-red light (730 mμ); the latter form can be converted back again to the original red absorbing form. These conversions have dramatic consequences; for example, red light inhibits stem elongation and lateral root formation but stimulates leaf expansion, chloroplast development, red flower coloration, and spore germination. Cycles of red and far-red light also can affect flower formation.

Indirect effects of light on plants

The effects of light on animals, although less obvious, may be important, as, for example, the effect of light on growth of the reproductive system of some animals. Increase in day length, hence in the amount of light, seems to initiate growth and development of the sex organs (gonads) in some birds during the spring. Curiously, the eyes are not the receptors for the light signal that activates the endocrine system to initiate growth of gonads; rather, cells deep in the brain are sensitive to the small amounts of light that pass directly through the thin skull of the bird.

Most animals show cyclic activity, or rhythms, in various important physical (*e.g.*, movement) and chemical (*e.g.*, respiration) events that are essential to the individual. These rhythms are often regulated by short exposure to light.

Chemical factors. Chemical factors of importance in the environment include the gases in the atmosphere and the water, mineral, and nutritional content of food. Plants require carbon dioxide, water, and sunlight for photosynthesis; drought slows plant growth and may even kill the plant. The effects of atmospheric contaminants—*e.g.*, oxides of nitrogen, hydrocarbons, and carbon monoxide—

are known to have deleterious effects on the growth and reproduction of both plants and animals.

Plants and animals require minerals and small amounts of elements such as zinc, magnesium, and boron. Nitrogen and phosphorus are provided to plants as nitrates and phosphates in the soil. Inadequate quantities of any nutritional factor in the soil result in poor plant growth and poor crop yields. Animals require oxygen, water, and elements from the environment. Because they are unable to synthesize sugars from carbon dioxide, animals must acquire these nutrients through the diet, either directly, by the consumption of plants, or indirectly, by the consumption of other animals that in turn have utilized plants as food. If the quality or quantity of this food is poor, either growth is retarded or death occurs (see NUTRITION).

Vitamins in diet

Vitamins, a class of compounds with a variety of chemical structures, are needed by animals in small amounts. Animals cannot synthesize all vitamins they require; those that cannot be synthesized must therefore be acquired in the diet, either from plants or from other animals that can synthesize the vitamin. Because certain vitamins are necessary in certain important metabolic reactions, vitamin deficiency during growth may have a variety of effects—stunting, malformation, disease, or death (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Vitamins*).

INTERNAL FACTORS

The organism is dependent on the environment for the raw materials for growth, but growth is also regulated internally. Because the size and form of plants and animals are under genetic control, events such as the rate and site of cell division and the extent of cell enlargement can be affected by mutations. It is not yet known, however, precisely how these factors, which are the ultimate determinants of growth, are controlled in individual cells.

One very important class of intrinsic growth regulators is that of the hormones. The principal plant hormone, auxin, is produced in the leaves; it moves by precise mechanisms, as yet poorly understood, to the other parts of the plant, controlling such processes as elongation of plant cells. Auxin somehow changes the characteristics of the rigid cell wall of the plant cell so that it becomes more flexible; the internal pressure within the cell then forces it to become larger. Other plant hormones may also play a role in the process; hormones such as cytokinins and gibberellins influence the rate of cell division in the meristems. Some dwarf plants can be stimulated to grow to normal size simply by applying gibberellin.

Hormones also play a decisive role in animal growth. One hormone from the pituitary gland at the base of the brain is called growth hormone because of its extensive and widespread effects on growth. A deficiency of growth hormone in pre-adolescents results in dwarfism, and oversupply of the hormone (often caused by a tumour) results in gigantism. If an excess of growth hormone is produced after the long bones can no longer grow—*i.e.*, post-adolescence—a disease called acromegaly, which is characterized by increases in the size of the hands and feet and broadening of facial features, results. A deficiency of thyroid hormone in children also causes growth retardation.

The sex hormones secreted from the pituitary gland interact in a complex way to regulate the growth of the gonads. The gonads in turn produce estrogen and progesterone in females and testosterone in males; these hormones control the development of human secondary sexual characteristics—body hair, enlargement of mammary glands in females, and growth of the vocal cords in males. Although the growth hormones and sex hormones play a vital role in growth, the exact mechanism by which they function has not been established with certainty (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Hormones*).

In addition to having the ability to synthesize the factors that regulate growth, plants and animals evidently possess exquisite mechanisms for integrating and regulating the production of hormones; *i.e.*, the appropriate amounts of the right hormones are produced at the right time and the right place for normal growth.

Limits of growth

Although many plants, including trees, grow throughout their lives, growth of parts of the organism is not perpetual; *e.g.*, leaves of a given species attain a specific size and can grow no larger. In animals, growth stops entirely, except for replacement, after the juvenile period. The limits for both total body size and organ size are probably established by genetic mechanisms. The factors involved in limiting the growth of an organism are not yet definitely known, but evidence indicates that the liver releases into the bloodstream protein molecules that can limit growth of the organ. Thus, one theoretical view is that an organ may produce substances that serve to limit its own growth, thereby establishing a feedback mechanism. A protein called nerve-growth factor is important for the growth of some parts of the mammalian nervous system. If too much of the nerve-growth factor is present, growth of sympathetic nerve fibres is extensive and aberrant. If the nerve-growth factor is eliminated from the body—by injection of an antibody against the factor—the sympathetic nerves wither and disappear. Other subtle growth regulatory substances specific for various organ systems may eventually be discovered.

The dynamics of growth

MEASUREMENT OF GROWTH

The mathematical analysis of the rate of growth has been a subject of interest for many years. It is based on the rule of cell division: one cell gives rise to two daughter cells. Hence, the theoretical increase in cell number would be a geometric series, in which one cell produces two cells, then four, eight, 16, and so on. In reality, however, the rate of growth is not constant but declines after a period of time, usually because of influences in the environment or because of inherent genetic limitations. Thus the curve showing the growth of cell populations and of organisms is usually S-shaped, or sigmoid, when growth is plotted against time on a graph (see Figure 1). The increase in

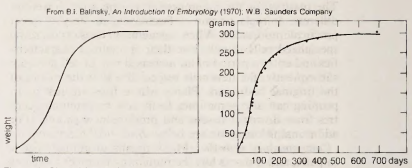


Figure 1: *Sigmoid growth curves.*
(Left) Ideal sigmoid growth curve. (Right) An actual postnatal growth curve of the white rat.

cell number resulting from cell division accounts for the rising part of the curve; the rate of cell division decreases at the plateau in the curve. The S-shaped growth curve is generally applicable to the growth of organisms. If growth is plotted against time on a logarithmic scale, the early intense growth (called log growth) in the rising phase of the growth curve falls on a straight line.

Rate of growth equation

The rate of growth may be defined by the differential equation $v = dW/dt\ (1/W)$, in which v is the growth rate and W is the weight at any given time, t. The solution of this equation provides a value for relative increase—the increase in weight related to the initial mass of the growing substance. The animal that most closely approaches a constant rate of growth is an insect larva. In most animals the rate of growth declines as the organism becomes larger and older.

Although the S-shaped growth curve describes with fair accuracy the growth of populations of single cells, such as bacteria or cells of higher organisms in tissue culture—the growth in a sterile nutrient environment of cells of tissues from organisms—the growth rates of different parts of whole organisms vary. The relationship of the growth of one part of an organism to that in another part is called allometry. An equation expressing the fundamental relationship of allometric growth is $y = bx^k$ in which y is the size of one organ; x is the size of another; b is a constant; and k is known as the growth ratio. Although such mathematical tools have allowed a very thorough

description of the differential growth of different parts of an organism, they have unfortunately not provided insight into the physical and chemical control of the growth rate.

THE STUDY OF GROWTH

Even though the chemical, physical, and genetic bases of growth are elusive, much has been learned about the process by growing tissues in a sterile nutrient environment. Even if the source of the tissue is an organ that has completely stopped growing, such as the nervous system of an animal or the phloem of a plant, the cells will begin to grow again in culture, often at a logarithmic rate of increase. It may therefore be concluded that the organism as a whole places constraints upon the ability of individual cells to reproduce and that, when these constraints are removed, the growth potential of the cells is no longer restrained. Even in tissue culture, however, the rate of cell growth eventually slows, hence the sigmoid-shaped growth curve. During the rapid growth phase of cells in tissue culture, they usually lose the ability to carry out the specialized function characteristic of their organ of origin; for example, if cartilage cells divide rapidly, they no longer synthesize cartilaginous matrix. This phenomenon of apparent despecialization has been a topic of great theoretical interest: are rapid growth and specialization mutually exclusive activities? Evidence shows that some types of specialized cells may be maintained in tissue culture for very long periods of time and still retain the ability to carry out specialized biosyntheses, so that the apparent loss of specialized function in tissue culture cells may not fundamentally result from a mutual exclusivity of growth and differentiation.

Contact inhibition of growth

When the growth of tissue-culture cells begins to slow, one factor responsible is exhaustion of critical components from the medium. But even if the medium is frequently replaced, when the bottom of the culture dish becomes densely packed with a layer of cells, the growth rate drops—a phenomenon called contact inhibition of growth. It is believed that cells so close that they are always touching provide a signal that retards the rate of cell division. Apparently identical cells in tissue culture also show great variation in growth rate. Some cells from the skin, for instance, when placed in culture, may divide every eight hours; other similar cells may divide only every 36 hours. The growth of cells in a controlled environment such as tissue culture offers many possibilities for studying the fundamental mechanisms controlling cell growth and, consequently, the growth of organisms and populations. (F.H.W.)

Aberrations of growth: biological malformation

The processes of development are regulated in such a way that few malformed organisms are found. Those that do appear may, when properly studied, shed light on normal development. The science of teratology—a branch of morphology or embryology—is concerned with the study of these structural deviations from the normal, whether in animals or plants.

In general, abnormalities can be traced to deviations from the normal course of development, often in very early embryonic stages. Such deviations may be caused by abnormal (mutant) genes, by environmental conditions, by infection, by drugs, and, perhaps most frequently, by interactions between these sets of causes. A general interpretation has been that one factor in many cases is reduction of the rate of development, the kind and degree of deformity depending upon the stage at which the retardation occurs. This interpretation is supported by the results of descriptive studies of anomalies, and especially by evidence from experimental teratology. (L.C.D.)

PLANT MALFORMATIONS

Monstrosities, freaks, and other malformations have interested botanists for many years. There are numerous categories of such growth abnormalities in plants, and these are often related only loosely or not at all to one another.

Exaggerated growth. Sometimes divergence from the normal represents merely a quantitative change, which is evidenced by a harmonious but exaggerated manifestation of the normal developmental processes. This is well illustrated in the so-called *bakanae*, or foolish seedling disease, of rice. The *bakanae* disease is caused by the fungus *Gibberella fujikuroi*. Diseased plants are often conspicuous in a field because of their extreme height and pale, spindly appearance. This exaggerated growth response was found to be due to specific substances, known as gibberellins, which were produced by the fungus. Evidence is now available to indicate that gibberellins, also produced by higher plant species, participate directly as an essential growth-regulating system in all higher plant species. The gibberellins of either fungal or higher plant origin stimulate the normal development of certain genetic dwarfs of maize and peas, which cannot themselves produce the gibberellins in amounts sufficient for their normal development.

Effect of the gibberellins

A common deformity of tobacco, called frenching, occurs in most tobacco-growing regions of the world. The advanced state of this condition is characterized by a cessation of terminal bud and stem growth. When dominance of the stem tips is lost, the buds in the axils of the leaves develop, and unusually large numbers of leaves (as many as 300) appear on a plant. The leaves are characteristically sword- or string-shaped because of the failure of the leaf blades to develop. Such plants have the apperance of a rosette. Although the cause of frenching has not yet been unequivocally established, it is thought to be due to a toxic substance produced by the nonpathogenic soil bacterium *Bacillus cereus*.

Alteration of floral parts. Under the stimulus of pathogenic organisms of the most diverse kinds, the sepals, petals, stamens, or pistils of a flower may be transformed into structures that are very different in appearance from those found normally. Certain viruses can cause enlargement of the leaflike flower parts (sepals) surrounding the base of a blossom in plants of the nightshade family. The tomato big-bud virus appears to affect the sepals of the tomato flower rather specifically. These structures enlarge greatly under the influence of the virus and fuse to form huge bladderlike structures that may be 10 times or more the normal size. In the Madagascar periwinkle (*Vinca rosea*), however, viruses of this type bring about a green colouring in the petals, stamens, and styles; normally the petals are pink and the stamens and styles whitish. There is in this instance a retrograde development of floral parts into foliage leaves. (Findings such as these are of interest to the morphologist because they support the contention that the flower should be regarded as a modified leafy branch.)

Translocation of organs. Plant organs may arise in unusual places as a result of the infection by certain types of pathogenic agents. The carrot-yellows virus, for example, stimulates production of aerial tubers in the axils of the leaves of potato plants. Large numbers of adventitious roots (arising in abnormal places) appear on the stems of tomato plants infected with the bacteria *Pseudomonas solanacearum* and *Agrobacterium tumefaciens* as well as the *Fusarium* wilt fungus and the cranberry false blossom virus.

An extreme example of adventitious shoot formation is found in *Begonia phyllomaniaca* after shock. In this instance, small plantlets develop spontaneously in incredible numbers from the superficial cell layers of the leaf blades, petioles, and stems. The adventitious shoots do not arise from preformed buds but develop from cells at the base of hairs and especially from certain glands present in great numbers in young stems and leaves of this species. Although these plantlets develop a vascular system of their own, the vast majority never succeed in connecting that system with the vascular system of the host. They must therefore be regarded not as branches but rather as independent organisms.

The independence of abnormal growths

Witches'-brooms. Witches'-brooms, or hexenbesens, are closely grouped, many-branched structures commonly found on a number of species of trees and shrubs and caused by certain fungi. Witches'-brooms live a more or less independent existence, despite the fact that they are derived from the tissues of the host. In accordance with their independence, the witches'-brooms tend to break

away from the normal correlations of the parent plant. Instead of branching out horizontally, the brooms stand as more or less erect clusters of branches. Witches'-brooms do not as a rule flower, and the vegetative buds may open several weeks earlier in the spring than do those present on healthy branches, indicating further the independence of these structures from normal controls.

Similar structures occur in certain plant species after virus infection. These appear to result from the excessive stimulation and development of secondary shoots. The witches'-broom virus in potatoes, for example, causes the infected plant to produce numerous buds on the above-ground stems of potato plants. Long, slender stolons resembling aerial roots that are covered with hairs develop from these adventitious buds.

Fasciation. This condition is best placed in that category of teratological abnormalities known as monstrosities. Fasciation is a term that has been used to describe a series of abnormal growth phenomena resulting from many different causes, all of which result in flattening of the main axis of the plant. Although a ribbonlike expansion of the stem is often the most striking feature of this condition, all parts of the plant may be affected. As fasciation develops, the growing point of the plant becomes broader; the unregulated tissue growth results in significant increases in the weight and volume of the plant. The apical growing point becomes linear and comblike in some instances or develops numerous growing points, producing a witches'-broom effect. In still other instances, the growing points may be coiled and resemble a ram's horn, or they may be fused and highly distorted into a grotesque tangle of coils. Fasciations found in plants such as the common cockscomb (*Celosia argentea cristata*) and in cacti are highly prized by gardeners.

There seems little doubt that nutritional changes due to disturbances in the growth-hormone relationships in a plant play an important role in fasciation. It has been suggested that maldistribution of growth hormones in the plant is also a cause of these abnormalities. (A.C.Br.)

ANIMAL MALFORMATIONS

Among the newborn young and embryos of man and most other species of animals are found occasional individuals who are malformed in whole or in part. The most grossly abnormal of these have been referred to from ancient times as monsters, probably because the birth of one was thought to signify something monstrous or portentous; the less severely malformed are usually known as abnormalities or anomalies.

The origin of malformations

Monsters have been regarded by primitive peoples as of supernatural origin. The birth of a malformed individual was often attributed, before the rise of modern science, to intercourse between human beings and devils or animals. The mythical beings that appear in the folklore of many peoples—races of dwarfs and giants, of sirens, mermaids, and men with a single median eye (cyclops) or leg (skiapods)—were probably suggested by observations of malformed humans. Giants and dwarfs were often classed as monsters, probably because of the prominent places they occupied in mythology.

The objective study of malformations began with the English physiologist William Harvey (1651), who correctly attributed them to deviations from the normal course of embryonic development. Systematic scientific study, however, had to await the pioneer work of the French anatomists Étienne and Isidore Geoffroy Saint-Hilaire. Their *Traité de Teratologie* (1836), which laid the basis for the science of teratology, still remains a valuable source of information. Recent improvements in understanding have come from the application of experimental analytical methods and from increased knowledge of the mechanisms of inheritance—*e.g.*, from genetics.

In man certain gross defects in babies at birth have been shown to be associated with effects acting through the mother in early pregnancy: gross defects of eyes and ears caused by infection of the mother with rubella (German measles); and microcephalic idiocy with diagnostic use of X ray on the mother. The latter observations have been confirmed by animal experimentation. It has been established that higher rates of congenital malformations occur in areas of higher natural radioactivity. The effect may be induced by fallout from atomic explosions.

The teratogenic action of many drugs has been tragically dramatized. The appearance of an alarming number of deformed, basically limbless infants, especially in Germany, in the late 1950s and early 1960s, was traced to the ingestion by pregnant women of the sedative thalidomide (known under many trade names). This drug adversely influences the developing fetus; it appears to interfere with development only in the first seven weeks of pregnancy. Hallucinogenic drugs, such as LSD (lysergic acid diethylamide), are suspected of damaging chromosomes, and their use could result in defective offspring.

According to form, two main classes of malformations may be recognized: those with defective or excessive growth in a single body, and those with partial or complete doubling of the body on one of its axes.

Repetition or deficiency of parts. *Somatic characters.* Repetition or deficiency of single parts, such as fingers or toes (polydactyly, hypodactyly [ectrodactyly], brachydactyly), is a frequent anomaly in man and other mammals. In many analyzed cases it has been shown to result from the inheritance of an abnormal gene that produces a localized disturbance of a growth process in the embryo. In the rabbit a recessive gene for brachydactyly (short digits) causes a localized breakdown of circulation in the developing limb bud of the embryo, followed by necrosis (tissue death) and healing.

Absence or abnormality of whole limbs is less common and includes, besides clubfoot, the so-called congenital amputations once thought to be caused by the strangulation of a limb by a fold of embryonic membrane (amnion). It is probable that internal abnormalities of the bone are more frequent causes of such amputations than are strangulations. Cases are recorded of human identical twins in which both members have the same type of limb abnormality, suggesting a hereditary predisposition to this type of malformation. Besides malformed individuals with rudimentary limbs (phocomelus; having "seal-like limbs"), others have incomplete or underdeveloped extremities (hemimelus, micromelus, ectromelus).

A rare type of malformation, but one that has always attracted special interest, occurs when the lower extremities are more or less united, as in the mythical figures of sirens or mermaids. Such sirenoid individuals may have a single foot (uromelus), or limbs fused throughout their length with no separate feet (sirenomelus or symmelus).

Defects of the brain and head of man

Absence of the brain at birth (anencephaly); an abnormally small brain and head (microcephaly); and enlargement of the brain and head, sometimes to prodigious dimensions due to dilation of the ventricles by fluid (hydrocephaly), are frequent congenital defects in man. In some cases they have been traced to defective genes, although they may also arise from accidental or traumatic processes during embryonic development. Occasionally, malformed persons are found in which a part of the brain protrudes through the cranium as an encephalocoele. An extreme variant of this type is pseudencephaly, in which the whole brain is everted and rests upon the top of the cranium like a wig.

Cyclopian malformations with a single median eye occur rarely in man and other animals. More frequent anomalies are anophthalmia (absence of eyes) and microphthalmia (abnormally small eyes), both occasionally the result of abnormal heredity. Defective closure of lines of junction in the embryo produces malformations such as cleft palate, in which the ventral laminae of the palate have failed to fuse, and harelip, in which the median nasal and maxillary processes fail to unite. A frequent abnormality in human infants is spina bifida, in which the spine fails to close over and a gap is left in the vertebral column. These conditions are inherited, albeit somewhat irregularly, in man.

Sexual anomalies. In man and other vertebrates, male and female individuals usually have distinctive characters in addition to the primary one of producing either sperm or eggs. Individuals with both male and female functions are known as hermaphrodites. While this is the normal condition in some lower animals and in many flowering

plants, it is so rare in mammals as to be regarded as anomalous. Individuals with mixtures of male and female characters (usually sterile) are known as intersexes. In man there occur two rare conditions that, according to recent evidence, represent partial sex reversal. Individuals with Klinefelter's syndrome are apparent males who produce no sperm. Many cases have been shown to have two X-chromosomes (the usual state determining femaleness) with an additional Y-chromosome (which carries genetic factors for maleness). Individuals with Turner's syndrome are apparent females without functional ovaries. The cases analyzed have only one X-chromosome (like the normal male with one X- and one Y-chromosome). These anomalies are clearly caused by disturbances in the mechanism for sex determination.

Complex syndromes. A remarkable feature of malformations in vertebrates including man is the association of multiple abnormalities in complex syndromes. Thus, in man harelip, spina bifida, hydrocephalus, and polydactyly may be found in the same individual; acrocephalosyndactyly (an egg- or dome-shaped skull and partial or complete fusion of digits in both hands and feet) often occurs with harelip, contractures, spina bifida, and mental abnormalities.

In man, individuals afflicted with mongolism, also known as Down's syndrome, have facial and bodily characters that permit diagnosis at or even before birth. Mongols have 47 instead of the normal 46 chromosomes. The extra chromosome is apparently responsible for the abnormal condition.

Doubling of parts. Individuals partially or wholly double, but joined together, are represented by the rare occurrence in man of Siamese twins, so-called from a famous Siamese pair exhibited for many years in the 19th century. The condition consists of identical twins joined by a bridge of tissue through which the circulatory systems communicate. Such twins probably arise by the incomplete separation of a single fertilized egg into two parts; the experimental production of such double individuals in newts has been accomplished by constricting the egg in the two-cell stage.

Siamese twins

In man, partially double symmetrical malformations are found. They vary from those with a single head but with neck, trunk, and limbs doubled, through those with two heads and a single trunk, to others with head, shoulders, and arms doubled, but with one trunk and one pair of legs. Such double malformations probably arise following the less complete separation of the halves of the early embryo or partial separation at later stages. A rare type is one in which there is a Janus head, two faces on a single head and body. Janus malformations have been produced experimentally in amphibian embryos by a variety of treatments in early stages. A group of cases in which the hinder end of the body was doubled from the sacrum back has been found in one strain of mice and appears to be due to abnormal heredity. Doubling of whole limbs in amphibia has been produced experimentally by injuring the limb rudiment at an early, sensitive stage. (For additional information on abnormal human development, see below *Human growth and development: abnormal development.*) (L.C.D.)

Biological regeneration

Organisms differ markedly in their ability to replace lost or amputated body parts. Some grow a new structure on the stump of the old one. By such regeneration whole organisms may dramatically replace substantial portions of themselves when they have been cut in two, or may grow organs or appendages that have been lost. Not all living things regenerate parts in this manner, however. The stump of an amputated structure may simply heal over without replacement. This wound healing is itself a kind of regeneration at the tissue level of organization: a cut surface heals over, a bone fracture knits, and cells replace themselves as the need arises.

Regeneration, as one aspect of the general process of growth, is a primary attribute of all living systems. Without it there could be no life, for the very maintenance of an organism depends upon the incessant turnover by which all tissues and organs constantly renew themselves. In some cases rather substantial quantities of tissues are replaced from time to time, as in the successive production of follicles in the ovary or the molting and replacement of hairs and feathers. More commonly, the turnover is expressed at the cellular level. In mammalian skin the epidermal cells produced in the basal layer may take several weeks to reach the outer surface and be sloughed off. In the lining of the intestines, the life span of an individual epithelial cell may be only a few days.

Cellular renewal

The motile, hairlike cilia and flagella of single-celled organisms are capable of regenerating themselves within an hour or two after amputation. Even in nerve cells, which cannot divide, there is an endless flow of cytoplasm from the cell body out into the nerve fibres themselves. New molecules are continuously being generated and degraded with turnover times measured in minutes or hours in the case of some enzymes, or several weeks as in the case of muscle proteins. (Evidently, the only molecule exempt from this inexorable turnover is deoxyribonucleic acid [DNA] which ultimately governs all life processes.)

There is a close correlation between regeneration and generation. The methods by which organisms reproduce themselves have much in common with regenerative processes. Vegetative reproduction, which occurs commonly in plants and occasionally in lower animals, is a process by which whole new organisms may be produced from fractions of parent organisms; *e.g.,* when a new plant develops from a cut portion of another plant, or when certain worms reproduce by splitting in two, each half then growing what was left behind. More commonly, of course, reproduction is achieved sexually by the union of an egg and sperm. Here is a case in which an entire organism develops from a single cell, the fertilized egg, or zygote. This remarkable event, which occurs in all organisms that reproduce sexually, testifies to the universality of regenerative processes. During the course of evolution the regenerative potential has not changed, but only the levels of organization at which it is expressed. If regeneration is an adaptive trait, it would be expected to occur more commonly among organisms that appear to have the greatest need of such a capability, either because the hazard of injury is great or the benefit to be gained is great. The actual distribution of regeneration among living things, however, seems at first glance to be a rather fortuitous one. It is difficult indeed to understand why some flatworms are able to regenerate heads and tails from any level of amputation, while other species can regenerate in only one direction or are unable to regenerate at all. Why do leeches fail to regenerate, while their close relatives, the earthworms, are so facile at replacing lost parts? Certain species of insects regularly grow back missing legs, but many others are totally lacking in this capacity. Virtually all modern bony fishes can regenerate amputated fins, but the cartilaginous fishes (including the sharks and rays) are unable to do so. Among the amphibians, salamanders regularly regenerate their legs, which are not very useful for movement in their aquatic environment, while frogs and toads, which are so much more dependent on their legs, are nevertheless unable to replace them. If natural selection operates on the principle of efficiency, then it is difficult to explain these many inconsistencies.

Whether regeneration is adaptive

Some cases are so clearly adaptive that there have evolved not only mechanisms for regeneration, but mechanisms for self-amputation, as if to exploit the regenerative capability. The process of losing a body part spontaneously is called autotomy. The division of a protozoan into two cells and the splitting of a worm into two halves may be regarded as cases of autotomy. Some colonial marine animals called hydroids shed their upper portions periodically. Many insects and crustaceans will spontaneously drop a leg or claw if it is pinched or injured. Lizards are famous for their ability to release their tails. Even the shedding of antlers by deer may be classified as an example of autotomy. In all these cases autotomy occurs at a predetermined point of breakage. It would seem that wherever nature contrives to lose a part voluntarily, it provides the capacity for replacement.

Compensatory hypertrophy

Sometimes, when part of a given tissue or organ is removed, no attempt is made to regenerate the lost structures. Instead, that which remains behind grows larger. Like regeneration, this phenomenon—known as compensatory hypertrophy—can take place only if some portion of the original structure is left to react to the loss. If three-quarters of the human liver is removed, for example, the remaining fraction enlarges to a mass equivalent to the original organ. The missing lobes of the liver are not themselves replaced, but the residual ones grow as large as necessary in order to restore the original function of the organ. Other mammalian organs exhibit similar reactions. The kidney, pancreas, thyroid, adrenal glands, gonads, and lungs compensate in varying degrees for reductions in mass by enlargement of the remaining parts.

It is not invariably necessary for the regenerating tissue to be derived from a remnant of the original tissue. Through a process called metaplasia, one tissue can be converted to another. In the case of lens regeneration in certain amphibians, in response to the loss of the original lens from the eye, a new lens develops from the tissues at the edge of the iris on the upper margin of the pupil. These cells of the iris, which normally contain pigment granules, lose their colour, proliferate rapidly, and collect into a spherical mass which differentiates into a new lens.

MODES OF REGENERATION

Basic patterns. Not all organisms regenerate in the same way. In plants and in coelenterates such as the hydra and jellyfishes, missing parts are replaced by reorganization of preexisting ones. The wound is healed, and the neighbouring tissues reorganize themselves into whatever parts may have been cut off. This process of reorganization, called morphallaxis, is the most efficient way for simple organisms to regenerate. Higher animals, with more complex bodies, regenerate parts differently, usually by the production of a specialized bud, or blastema, at the site of amputation. The blastema, made up of cells that look very much alike despite their often diverse origins, made its first appearance evolutionarily in flatworms and is encountered in the regenerative processes of all higher animals. It provides the tissue that will form the regenerated part.

Atypical regeneration. Sometimes the part that grows back is not the same as that which was lost, and, occasionally, regeneration may be induced without having lost anything at all. It is not uncommon for a regenerated part to be incomplete. Earthworms, for example, usually regenerate only five segments in the anterior direction even if more than that number have been amputated. Many insects regenerate abnormally small legs from which some segments may be missing. Tadpole tails when amputated grow back to about only half their original length. These and other cases testify to the fact that a little regeneration is often good enough—that it is not necessary in every case to reproduce a flawless copy of the original.

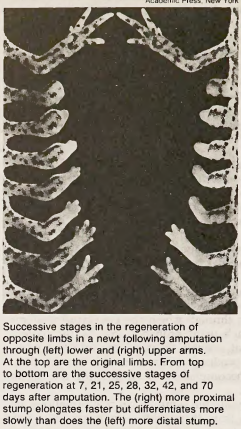
Sometimes that which is regenerated is very different from the original. Among the arthropods there are cases in which the stump of an antenna grows a leg, while a cut eyestalk regenerates an antenna. More commonly, the regenerated part may be a reasonable facsimile of the original but will differ in details. A regenerated lizard tail contains an unsegmented cartilaginous tube instead of a series of vertebrae as did the original tail. The spinal cord lacks segmented ganglia, and the scales in the skin differ in character from the original ones. A regenerated tail, therefore, is easily distinguished from an original one yet appears sufficient to serve the purpose. Another interesting case is that of jaw regeneration in salamanders. If the lower jaw is amputated a new one will grow back, but it is often smaller than the original. It contains teeth and a mandible, but lacks a new tongue. Furthermore, the new mandible is a cartilaginous model of the original, and is not known to convert into bone.

Sometimes more of a part grows back than has been removed by amputation. A limb stump, for example, can occasionally give rise to hands with extra digits. Lobsters have been known to regenerate double structures, in which case the new parts are mirror images of each other.

THE REGENERATION PROCESS

Origin of regeneration material. Following amputation, an appendage capable of regeneration develops a blastema from tissues in the stump just behind the level of amputation (see photograph). These tissues undergo drastic changes. Their cells, once specialized as muscle, bone, or cartilage, lose the characteristics by which they are normally identified (dedifferentiation); they then begin to migrate toward, and accumulate beneath, the wound epidermis, forming a rounded bud (blastema) that bulges out from the stump. Cells nearest the tip of the bud continue

R.J. Goss, *Principles of Regeneration* (1969), Academic Press, New York



Successive stages in the regeneration of opposite limbs in a newt following amputation through (left) lower and (right) upper arms. At the top are the original limbs. From top to bottom are the successive stages of regeneration at 7, 21, 25, 28, 32, 42, and 70 days after amputation. The (right) more proximal stump elongates faster but differentiates more slowly than does the (left) more distal stump.

to multiply, while those situated closest to the old tissues of the stump differentiate into muscle or cartilage, depending upon their location. Development continues until the final structures at the tip of the regenerated appendage are differentiated, and all the proliferating cells are used up in the process.

The blastema cells seem to differentiate into the same kind of cells they were before, or into closely related types. Cells may perhaps change their roles under certain conditions, but apparently rarely do so. If a limb blastema is transplanted to the back of the same animal, it may continue its development into a limb. Similarly, a tail blastema transplanted elsewhere on the body will become a tail. Thus, the cells of a blastema seem to bear the indelible stamp of the appendage from which they were produced and into which they are destined to develop. If a tail blastema is transplanted to the stump of a limb, however, the structure that regenerates will be a composite of the two appendages.

Polarity and gradient theory. Each living thing exhibits polarity, one example of which is the differentiation of an organism into a head, or forward part, and a tail, or hind part. Regenerating parts are no exception; they exhibit polarity by always growing in a distal direction (away from the main part of the body). Among the lower invertebrates, however, the distinction between proximal (near, or toward the body) and distal is not always clear cut. It is not difficult, for example, to reverse the polarity of "stems" in colonial hydroids. Normally a piece of the stem will grow a head end, or hydranth, at its free, or distal, end; if that is tied off, however, it regenerates a hydranth at the end that was originally proximal. The polarity in this system is apparently determined by an activity gradient in such a way that a hydranth regenerates wherever the metabolic rate is highest. Once a hydranth

has begun to develop, it inhibits the production of others proximal to it by the diffusion of an inhibitory substance downward along the stem.

Differences between whole organisms and appendages

When planarian flatworms are cut in half, each piece grows back the end that is missing. Cells in essentially identical regions of the body where the cut was made form blastemas, which, in one case gives rise to a head and in the other becomes a tail. What each blastema regenerates depends entirely on whether it is on a front piece or a hind piece of flatworm: the real difference between the two pieces may be established by metabolic differentials. If a transverse piece of a flatworm is cut very thin—too narrow for an effective metabolic gradient to be set up—it may regenerate two heads, one at either end. If the metabolic activity at the anterior end of a flatworm is artificially reduced by exposure to certain drugs, then the former posterior end of the worm may develop a head.

Appendage regeneration poses a different problem from that of whole organisms. The fin of a fish and the limb of a salamander have proximal and distal ends. By various manipulations, it is possible to make them regenerate in a proximal direction, however. If a square hole is cut in the fin of a fish, regeneration takes place as expected from the inner margin, but may also occur from the distal edge. In the latter case, the regenerating fin is actually a distal structure except that it happens to be growing in a proximal direction.

Amphibian limbs react in a similar manner. It is possible to graft the hand of a newt to the nearby body wall, and once a sufficient blood flow has been established, to sever the arm between the shoulder and elbow. This creates two stumps, a short one consisting of part of the upper arm, and a longer one made up of the rest of the arm protruding in the wrong direction from the side of the animal. Both stumps regenerate the same thing, namely, everything normally lying distal to the level of amputation, regardless of which way the stump was facing. The reversed arm therefore regenerates a mirror image of itself.

Clearly, when a structure regenerates it can only produce parts that normally lie distal to the level of amputation. The participating cells contain information needed to develop everything "downstream," but can never become more proximal structures. Regeneration, like embryonic development, occurs in a definite sequence.

Regulation of regeneration. There are certain prerequisites without which regeneration cannot occur. First and foremost, there must be a wound, although the original appendage need not have been lost in the process. Second, there must be a source of blastema cells derived from remnants of the original structure or an associated one. Finally, regeneration must be stimulated by some external force. The stimuli often involve the nervous system. An adequate nerve supply is required for the regeneration of fish fins, taste barbels, and amphibian limbs. In the case of many tail regenerations, the spinal cord provides the necessary stimulus. Lens regeneration in salamander eyes depends upon the presence of a retina. Arthropod appendages regenerate in the presence of molting hormones. Protozoan regeneration requires the presence of a nucleus. In case after case, regeneration depends on more than a healed wound and a source of blastema cells. It is often triggered by some physiological stimulus originating elsewhere in the body, a stimulus invariably associated with the very function of the structure to be regenerated. The conclusion is inescapable that regeneration is primarily the recovery of deficient functions rather than simply the replacement of lost structures.

The imperative of need is of further importance in suppressing excess regeneration. To be able to regenerate is to run the risk of regenerating too much or too often. If regeneration did not depend upon a physiological stimulus, such as those mediated by nerves or hormones, there would be no reason why simple wounds should not sprout whole new appendages.

Failure of regenerative processes

It is not known why regeneration fails to occur in many cases, as in the legs of frogs or the limbs and tails of mammals. The nerve supply might be inadequate, for when the number of nerves is artificially increased, regeneration is sometimes induced. This cannot be the whole answer, however, because not all appendages depend on nerves for their regeneration; newt jaws, salamander gills, and deer antlers do not require nerves to regenerate.

Possibly the failure to regenerate relates to the ways in which wounds heal. In higher vertebrates there is a tendency to form thick scar tissue in healing wounds, which may act as a barrier between the epidermis and the underlying tissues of the stump. In the absence of direct contact between these two tissues, the stump may not be able to give rise to the blastema cells required for regeneration.

THE RANGE OF REGENERATIVE CAPABILITY

Virtually no group of organisms lacks the ability to regenerate something. This process, however, is developed to a remarkable degree in lower organisms, such as protists and plants, and even in many invertebrate animals such as earthworms and starfishes. Regeneration is much more restricted in higher organisms such as mammals, in which it is probably incompatible with the evolution of other body features of greater survival value to these complex animals.

Protists and plants. *Algae.* One of the most outstanding feats of regeneration occurs in the single-celled green alga *Acetabularia.* This plant-like protist of shallow tropical water consists of a group of short rootlike appendages; a long thin "stem," up to several centimetres in length; and an umbrella-like cap at the top. The entire organism is one cell, with its single nucleus situated at the base in one of the "roots." If the cap is cut off, a new one regenerates from the healed over stump of the amputated stem. The nucleus is necessary for this kind of regeneration, presumably because it provides the information needed to direct the development of the new cap. Once this information has been produced by the nucleus, however, the nucleus can be removed and regeneration continues unabated.

If the nucleus from one species of *Aceiabularia* is added to a cell-body of another species, and the cap of the recipient cell is amputated, the new cap that regenerates will be a hybrid because each nucleus exerts its own morphogenetic influences. On the other hand, if the nucleus from one species is substituted for that in another, regeneration reflects the properties of the new nucleus.

Protozoans. Most single-celled, animal-like protists regenerate very well. If part of the cell fluid, or cytoplasm, is removed from *Amoeba,* it is readily replaced. A similar process occurs in other protozoans, such as flagellates and ciliates. In each case, however, regeneration occurs only from that fragment of the cell containing the nucleus. Amputated parts that lack a nucleus cannot survive. In some ciliates, such as *Blepharisma* or *Stentor,* the nucleus may be elongated or shaped like a string of beads. If either of these organisms is cut in two so that each fragment retains part of the elongated nucleus, each half proceeds to grow back what it lacks, giving rise to a complete organism in less than six hours. The way in which such a bisected protozoan regenerates is almost identical with the way it reproduces by ordinary division. Even a very tiny fragment of the whole organism can regenerate itself, provided it contains some nuclear material to determine what is supposed to be regenerated.

Nuclear material

Green plants. The mechanisms by which vascular plants grow have much in common with regeneration. Their roots and shoots elongate by virtue of the cells in their meristems, the conical growth buds at the tip of each branch. These meristems are capable of indefinite growth, especially in perennial plants. If they are amputated they are not replaced, but other meristems along the stem, normally held in abeyance, begin to sprout into new branches that more than compensate for the loss of the original one. Such a process is called restitution.

Plants are also capable of producing callus tissue wherever they may be injured. This callus is proliferated from cambial cells, which lie beneath the surface of branches and are responsible for their increase in width. When a callus forms, some of its cells may organize into growing points, some of which in turn give rise to roots while others produce stems and leaves.

Callus tissue

Invertebrates. *Coelenterates.* The vast majority of research on coelenterates has been focussed on hydras and

some of the colonial hydroids. If a hydra is cut in half, the head end reconstitutes a new foot, while the basal portion regenerates a new hydranth with mouth and tentacles. This seemingly straightforward process is deceptively simple. From tiny fragments of the organism whole animals can be reconstituted. Even if a hydra is minced and the pieces scrambled, the fragments grow together and reorganize themselves into a complete whole. The indestructibility of the hydra may well be attributed to the fact that even the intact animal is constantly regenerating itself. Just below the mouth is a growth zone from which cells migrate into the tentacles and to the foot where they eventually die. Hence, the hydra is in a ceaseless state of turnover, with the loss of cells at the foot and at the tips of the tentacles being balanced by the production of new ones in the growth zone. If such an animal is X-rayed, the proliferation of new cells is inhibited and the hydra gradually shrinks and eventually dies owing to the inexorable demise of cells and the inability to replace them.

In colonial hydroids, such as *Tubularia,* there is a series of branching stems, each of which bears a hydranth on its end. If these hydranths are amputated they grow back within a few days. In fact, the organism normally sheds its hydranths from time to time and regenerates new ones naturally.

Planarian regeneration

Flatworms. Planarian flatworms are well-known for their ability to regenerate heads and tails from cut ends (see Figure 2). In the case of head regeneration, some

After (top) T.H. Morgan in E. Korschelt, *Regeneration und Transplantation,* from (bottom) V. Hamburger, *Manual of Experimental Embryology* (© 1960), University of Chicago Press

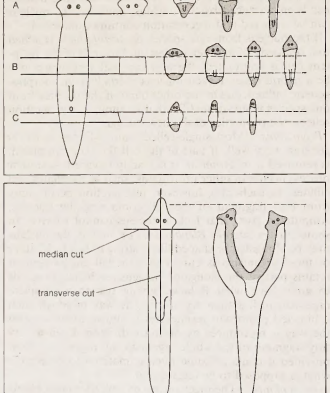


Figure 2: (Top) Regeneration in the planarian flatworm *Dugesia.* The three rows show regeneration from (A) anterior section, (B) midsection, and (C) posterior section of the animal at the left. (Bottom) Regeneration of a double head in a planarian. Regenerated tissue is shaded.

blastema cells become brain tissues, others develop into the eyes, and still others differentiate as muscle or intestine. In a week or so, the new head functions almost as well as the original.

The blastema that normally gives rise to a single head is, under certain circumstances, even capable of becoming two heads if the stump of a decapitated flatworm is divided in two by a longitudinal cut. Each of the two halves then gives rise to a complete head. Thus, each blastema develops into an entire structure regardless of its size or position in relation to the rest of the animal.

In the case of flatworms there is still considerable disagreement concerning the origins of the blastema. Some investigators contend that it is derived from neoblasts, undifferentiated reserve cells scattered throughout the body. Others claim that there are no such reserve cells and that the blastema develops from formerly specialized cells near the wound that dedifferentiate to give rise to the blastema cells. Whatever their source, the cells of the blastema are capable of becoming many different things depending upon their location.

Regeneration in flatworms occurs in a stepwise fashion. The first tissue to differentiate is the brain, which induces the development of eyes. Once the head has formed, it in turn stimulates the production of the pharynx. The latter then induces the development of reproductive organs farther back. Thus, each part is necessary for the successful development of those to come after it; conversely, each part inhibits the production of more of itself. If decapitated flatworms are exposed to extracts of heads, the regeneration of their own heads is prevented. Such a complex interplay of stimulators and inhibitors is responsible for the successful regeneration of an integrated morphological structure.

Annelids. The segmented worms exhibit variable degrees of regeneration. The leeches, as already noted, are wholly lacking in the ability to replace lost segments, whereas the earthworms and various marine annelids (polychaetes) can often regenerate forward and backward. The expression of such regenerative capacities depends very much on the level of amputation. Anteriorly directed regeneration usually occurs best from cuts made through the front end of the worm, with little or no growth taking place from progressively more posterior bisections. Posteriorly directed regeneration is generally more common and extensive. Some species of worms replace the same number of segments as were lost. Hypomeric regeneration, in which fewer segments are produced than were removed, is more common, however.

Anterior regeneration depends upon the presence of the central nerve cord. If this is cut or deflected from the wound surface, little or no forward regeneration may take place. Posterior regeneration requires the presence of the intestine, removal of which precludes the formation of hind segments. Thus, it would seem that no head will regenerate without a central nervous system, nor a tail without an opening.

Arthropods. Many insects and crustaceans regenerate legs, claws, or antennas with apparent ease. When insect legs regenerate, the new growth is not visible externally because it develops within the next proximal segment in the stump. Not until the following molt is it released from its confinement to unfold as a fully developed leg only slightly smaller than the original. In the case of crabs, regenerating legs bulge outward from the amputation stump. They are curled up within a cuticular sheath, not to be extended until the sheath is molted. Lobsters and crayfish regenerate claws and legs in a straightforward manner as direct outgrowths from the stumps. As in other crustaceans, however, these regenerates lie immobile within an enveloping cuticle and do not become functional until their sheath is shed at the next molt.

Molting

In all arthropods regeneration is associated with molting, and therefore takes place only during larval or young stages. Most insects do not initiate leg regeneration unless there remains ample time prior to the next scheduled molt for the new leg to complete its development. If amputation is performed too late in the intermolt period, the onset of regeneration is delayed until after shedding; the regenerate then does not appear until the second molt. Metamorphosis into the adult stage marks the end of molting in insects, and adults accordingly do not regenerate amputated appendages.

Crustaceans often tend to molt and grow throughout life. They therefore never lose the ability to grow back missing appendages. When a leg is lost, a new outgrowth appears even if the animal is not destined to molt for many months. Following a period of basal growth, during which a diminutive limb is produced, the regenerated part eventually ceases to elongate. Not until a few weeks before the

next molt does it resume growth and complete its development, triggered by the hormones that induce molting.

Vertebrates. *Fishes.* Many different parts of the fish's body will grow back. Plucked scales are promptly replaced by new ones, and amputated gill filaments can regenerate easily. The "whiskers," or taste barbels, of the catfish grow back as perfect replicas of the originals. The most conspicuous regenerating structures in fishes, however, are the fins. When any of these are amputated, new fins grow out from the stumps and soon restore everything that was missing. Even the coloured stripes or spots that adorn some fins are reconstituted by new pigment cells that repopulate the regenerated part. Fin regeneration depends on an adequate nerve supply. If the nerves are cut leading into the fin, regeneration of neither the amputated fin nor excised pieces of the bony fin rays can take place.

Fin regeneration

Amphibians. Salamanders are remarkable for their ability to regenerate limbs. Larval frogs, or tadpoles, also possess this ability, but usually lose it when they become frogs. It is not known why frog legs do not regenerate, and under appropriate stimuli they can be induced to do so.

Tadpoles and salamanders can replace amputated tails. Tadpole tails have a stiff rod called the notochord for support, whereas salamanders possess a backbone, composed of vertebrae. Both tails contain a spinal cord. When the salamander regenerates its tail, the spinal cord grows back and segmental nerve-cell clusters (ganglia) differentiate. Tadpoles also regenerate their spinal cords, but not the associated ganglia. If the spinal cord is removed or destroyed in the salamander, no tail regeneration occurs; if it is removed from the tadpole tail, however, regeneration can proceed without it.

Reptiles. Lizards also regenerate their tails, especially in those species that have evolved a mechanism for breaking off the original tail when it is grasped by an enemy. When the lizard tail regenerates, however, it does not replace the segmented vertebrae. Instead, there develops a long tapering cartilaginous tube within which the spinal cord is located and outside of which are segmented muscles. The spinal cord of the lizard tail is necessary for regeneration, but the regenerated tail does not reproduce the ganglia that are normally associated with it. Occasionally, a side tail may be produced if the original tail is broken but not lost.

Birds. Regeneration of amputated appendages in birds is not known to occur; however, they do replace their feathers as a matter of course. While most species shed and regenerate feathers one at a time so as not to be grounded, flightless birds, such as penguins, may molt them all at once. Male puffins cast off their colorful beaks after the mating season, but grow new ones the following year. In like manner, the dorsal keel on the upper beaks of male pelicans is shed and replaced annually.

Mammals. Although mammals are incapable of regenerating limbs and tails, there are a few exceptional cases in which lost tissues are in fact regenerated. Not the least of these cases is the annual replacement of antlers in deer. These remarkable structures, which normally grow on the heads of male deer, consist of an inner core of bone enveloped by a layer of skin and nourished by a copious blood supply. During the growing season the antlers elongate by the proliferation of tissues at their growing tips. The rate of growth in some of the larger species may surpass one centimetre (0.39 inch) per day; the maximum rate of growth recorded for the elk is 2.75 centimetres (1.05 inches) per day. When the antlers have reached their full extent, the blood supply is constricted, and the skin, or velvet, peels off, thus revealing the hard, dead, bony antlers produced by the male deer in time for the autumn mating season. The regeneration of elk antlers spans about seven months (see Figure 3). The following spring, the old antlers are shed and new ones grow to replace them.

Antler replacement

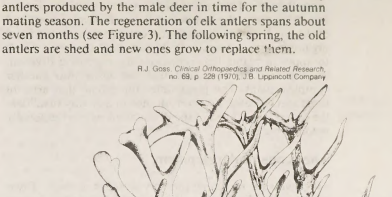


Figure 3: Regeneration of antlers in the elk (see text).

Still another example of mammalian regeneration occurs in the case of the rabbit's ear. When a hole is punched through the external ear of the rabbit, tissue grows in from around the edges until the original opening is reduced or obliterated altogether. This regeneration is achieved by the production of new skin and cartilage from the margins of the original hole. A similar phenomenon occurs in the case of the bat's wing membrane. (R.J.G.)

GENERAL FEATURES OF BIOLOGICAL DEVELOPMENT

Development in its most general meaning refers to any process of progressive change. In this sense, most modern philosophical outlooks would consider that development of some kind or other characterizes all things, in both the physical and biological worlds. Such points of view go back to the very earliest days of philosophy.

Among the pre-Socratic philosophers of Greek Ionia, half a millennium before Christ, some, like Heracleitus, believed that all natural things are constantly changing. In contrast, others, of whom Democritus is perhaps the prime example, suggested that the world is made up by the changing combinations of atoms, which themselves remain unaltered, not subject to change or development. The early period of post-Renaissance European science may be regarded as dominated by this latter atomistic view, which reached its fullest development in the period between Newton's laws of physics and Dalton's atomic theory of chemistry in the early 19th century. This outlook was never easily reconciled with the observations of biologists, and in the last hundred years a series of discoveries in the physical sciences have combined to swing opinion back toward the Heracleitan emphasis on the importance of process and development. The atom, which seemed so unalterable to Dalton, has proved to be divisible after all, and to maintain its identity only by processes of interaction between a number of component subatomic particles, which themselves must in certain aspects be regarded as processes rather than matter. Albert Einstein's theory of relativity showed that time and space are united in continuum, which implies that all things are involved in time; that is to say, in development.

Early theories of development

The philosophers who charted the transition from the nondevelopmental view, for which time was an accidental and inessential element, were Henri Bergson and, in particular, Alfred North Whitehead. Karl Marx and Friedrich Engels, with their insistence on the difference between dialectical and mechanical materialism, may be regarded as other important innovators of this trend, although the generality of their philosophy was somewhat compromised by the political context in which it was placed and the rigidity with which their later followers have interpreted it.

Philosophies of the Heracleitan type, which emphasize process and development, provide much more appropriate frameworks for biology than do philosophies of the atomistic kind. Living organisms confront biologists with changes of various kinds, all of which could be regarded

as in some sense developmental; however, biologists have found it convenient to distinguish the changes and to use the word development for only one of them. Biological development can be defined as the series of progressive, nonrepetitive changes that occur during the life history of an organism. The kernel of this definition is to contrast development with, on the one hand, the essentially repetitive chemical changes involved in the maintenance of the body, which constitute "metabolism," and on the other hand, with the longer term changes, which, while nonrepetitive, involve the sequence of several or many life histories, and which constitute evolution.

The blurring of distinctions in nature

As with most formal definitions, these distinctions cannot always be applied strictly to the real world. In the viruses, for instance, and even in bacteria, it is difficult to make a distinction between metabolism and development, since the metabolic activity of a virus particle consists of little more than the development of new virus particles. In certain other cases, the distinction between development and evolution becomes blurred: the concept of an individual organism with a definite life history may be very difficult to apply in plants that reproduce by vegetative division, the breaking off of a part that can grow into another complete plant. The possibilities for debate that arise in these special cases, however, do not in any way invalidate the general usefulness of the distinctions as conventionally made in biology.

The scope of development

All organisms, including the very simplest, consist of two components, distinguished by a German biologist, August Weismann, at the end of the 19th century, as the "germ plasm" and the "soma." The germ plasm consists of the essential elements, or genes, passed on from one generation to the next, and the soma consists of the body that may be produced as the organism develops (Figure 4). In more modern terms, Weismann's germ plasm is identified with DNA (deoxyribonucleic acid), which carries, encoded in the complex structure of its molecule, the instructions necessary for the synthesis of the other compounds of the organism and their assembly into the appropriate structures. It is this whole collection of other compounds (proteins, fats, carbohydrates, and others) and their arrangement as a metabolically functioning organism that constitutes the soma. Biological development encompasses, therefore, all the processes concerned with implementing the instructions contained in the DNA. Those instructions can only be carried out by an appropriate executive machinery, the first phase of which is provided by the cell that carries the DNA into the next generation: in animals and plants by the fertilized egg cell; in viruses by the cell infected. In life histories that have more than a minimal degree of complexity, the executive machinery itself becomes modified as the genetic instructions are gradually put into operation, and new mechanisms of protein synthesis are brought into functional condition. The fundamental problem of developmental biology is to understand the interplay between the genetic instructions and the mechanisms by which those instructions are carried out.

Genotype versus phenotype

In the language of genetics the word genotype is used to indicate the hereditary instructions passed on from one generation to another in the genes, while phenotype is the term given to the functioning organisms produced by those instructions. Biological development, therefore, consists of the production of phenotypes. The point made in the last paragraph is that the formation of the phenotype of one generation depends on the functioning of part of the phenotype of the previous generation (*e.g.*, egg cell), as the mechanism that begins the interpretation of the instructions contained in the new organism's genotype.

TYPES OF DEVELOPMENT

In the entire realm of organisms, many different modes of development are found, the most important categories of which can be discussed as pairs of contrasting types.

Quantitative and qualitative development. Development may amount to no more than a quantitative change (usually an increase) in a system that remains essentially unaltered. Qualitative development involves an alteration in the nature of the system. Pure examples of the first type are difficult to find. Approximations to it occur when an animal or plant has attained a structure with the full complement of organs; it then appears to increase only in size, that is to say, quantitatively. This would be a period of simple growth. A closer examination nearly always shows that the system is also undergoing some qualitative change, however. A human infant at birth, for example, already has its full complement of organs, but the ensuing developmental period up to adulthood involves not only growth but also processes of maturation that involve qualitative as well as quantitative changes. Perhaps the most uncomplicated examples of quantitative development occur in certain simple plants and animals. Flatworms, for example, may become reduced in size when starved but increase in size again when provided with suitable nutrition; they thus undergo quantitative changes. Even in these cases, however, it is found that the constituent organs do not always merely become reduced in size but may actually suffer the loss of certain parts.

Progressive and regressive development. The normal processes of development in the majority of plants and animals may be considered progressive since they lead to increases in size and complexity and to the addition of

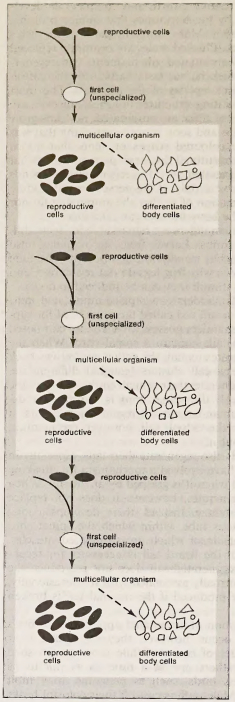


Figure 4: Weismann's concept of the continuity of the germ plasm.

From *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed. (1968); Rand McNally & Company

The adaptability of regression

new elements to the system. As already indicated, some organisms, when placed in adverse conditions, may undergo regressive changes, both in size and complexity. Such regressive changes are a part of the normal life history of certain organisms. Characteristically, these are species in which the organism at an early stage develops a relatively complex structure that enables it to be motile, and later adopts a form of life for which motility is no longer a necessity. A good example is that of the barnacles, a group of marine crustaceans in which the egg at first develops into a motile larva that soon settles down and becomes firmly attached to a solid underwater surface. The barnacle then loses many of the organs characteristic of the motile phase and develops into its familiar stationary form.

There are a number of other examples, particularly in groups in which the adults adopt a parasitic form of life, especially within the digestive system or other tissues of a host animal, from which they have only to absorb their nutriment without having to move or to possess suitable organs for capturing prey. In such cases the early developmental period is characterized by progression toward more complex forms followed by a period of regression in which many of these organs may be lost. During this regressive period certain components of the organism (*i.e.*, those concerned with functioning as a sessile or parasitic form) may undergo progressive development at the same time as the other organs are regressing.

Single-phase and multiphase development. The most familiar organisms, including man, undergo a single-phase development; the organs that appear at early stages persist throughout the whole of life. There are many kinds of animals that develop one or more larval stages adapted to a life different from that of the adult. Perhaps the best known of these is the common frog. The egg first develops into a tadpole, which is provided with a large muscular tail by which it swims. The tadpole eventually undergoes a change of form, or metamorphosis. This involves the regression and resorption of the tail and the growth of the limbs. During this time the rest of the body of the tadpole undergoes less profound changes; the organs persist but undergo relatively far-reaching progressive changes. In other animals, the alteration between the larval and the adult forms may be much more drastic. The egg of a sea urchin, for instance, at first develops to a small larva (the pluteus), which is completely unlike that of the adult. During metamorphosis nearly all the structures of the pluteus disappear; the five-rayed adult develops from a very small rudiment within the larva. In other groups of marine invertebrates, there may be successive larval stages before the adult form appears.

Alternation of generations

Plants in general appear to exhibit a type of development related in a general way to the multiphased development just discussed in animals, although rather different from it in essence. This is called the "alternation of generations" (Figure 5). The majority of higher plants possess two sets of similar chromosomes in each of their cells, that is to say they are diploid ($2n$), as are most higher animals. But in sexual reproduction, diploid cells undergo a reduction division so as to form precursors of the sex cells, which are haploid—*i.e.*, they contain only one set of chromosomes. In animals these cells develop directly into the sex cells—egg and sperm—which unite in fertilization. In plants the haploid cells undergo some developmental processes before the functioning sex cells are produced. The products of this development are spoken of as the "haploid generation." In most higher plants the haploid development is quite reduced, so that the haploid individuals contain only a few nuclei—those associated with the pollen tube on the male side and a few associated with the egg on the female side. In some lower plants, however, such as mosses and ferns, the haploid development may be much more extensive and give rise to quite sizable separate plants. In such cases a species contains two kinds of individuals, produced by different types of developmental processes controlled, however, by the same genotype. This may be compared with the multiphasing development of larval forms in animals. The situation in plants, however, is characterized by the two forms of the organism having different chromosomal constitutions—haploid and diploid—whereas the larval forms and the adult of an animal species have the same chromosomal constitution.

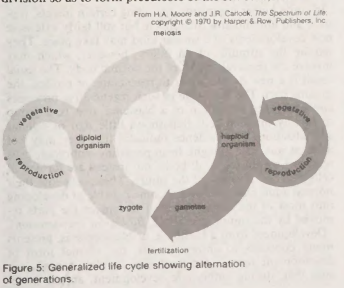


Figure 5: Generalized life cycle showing alternation of generations.

Structural and functional development. These two categories cannot be regarded as a pair of opposites as were the previous pairs in this list; rather, they are two aspects of all processes of biological development and can be separated only conceptually, and for purposes of convenience of description. Function is the capacity of the biological system to carry out operations. At the level of the organism, these operations include walking, swimming, eating, digesting, etc.; at the cell level, typical functions are respiring, contracting, conducting nervous impulses, secreting hormones, etc.; and at the molecular level, all functions depend on the production of enzymes, coded by particular genes. Structure encompasses all parts of the organism capable of carrying out functions localized within the body of the organism and arranged in some particular spatial pattern. Contractile cells, for example, are grouped together to form muscle, and other cells are grouped together to form elements of the skeleton; both the muscles and the skeletal elements have definite spatial relations to each other.

Form and function

These two aspects of development—function and structure—are not opposed to each other in any way. On the contrary, it is obvious that the higher level functions are clearly dependent on the proper structural relations and functions of cell systems. Even at the basic cellular or molecular levels, secretion or nervous conduction essentially depends upon the proper structural relation of the subcellular elements. It is, however, often convenient to focus discussion on one or other of these two aspects of development; for instance, a study may be made into the developmental processes that bring about the production of hemoglobin or insulin by a certain kind of cell, without at the moment being concerned with structural problems. Or again, the focus may be on the results of a certain process by which a mass of cells develops into a typical hand with five digits. In such an inquiry the structural aspects are paramount.

Normal and abnormal development. If a number of fertilized eggs of a given species are provided with conditions that enable them to develop at all, they will, with extraordinary regularity, develop into exceedingly similar adult organisms. The range of conditions they can tolerate is rather wide, and the similarity of the end products surprisingly complete. There are, indeed, good grounds for recognizing what must be considered normal development. The situation is perhaps more marked in animals than in plants, since the plants produced from a given batch of seed under a variety of environmental conditions often present considerably greater variation than is commonly found among animals. Even among plants, however, the differences produced by different conditions of cultivation are usually no more than quantitative differences in size and number of such organs as leaves and flowers, so that an individual can be described as well or poorly developed rather than as normally or abnormally developed. It is only in relatively few cases that a plant develops in quite different ways under two different conditions, neither of

The latitude of normal development

which can be considered abnormal or normal. In certain aquatic plants, for instance, the shape of the emergent leaves is different from the leaves that develop underwater. In such cases the plant actually has two normal forms of development.

It is possible, of course, to produce abnormal organisms by submitting a developing system to stimuli not usually encountered in a normal environment, such as certain chemicals. The presence of unusual genes also may result in deviations from the normal processes of development. In the vast majority of cases such abnormalities can be regarded as resulting from failure to carry out fully the normal processes of development. Functional abnormality in the adult consists in the failure of the system to produce a certain enzyme or functional cell type; a structural abnormality consists in the unusual appearance of certain component elements or in their arrangement in incompletely realized patterns. It is extremely rare to find examples in which the abnormality consists in the addition of a new enzyme not produced in normal development, or the formation of a new structural pattern of the elements.

One very important type of development that, from some points of view, can be considered as an exception to the rule that abnormal development is nearly always retrogressive, is carcinogenesis, the production of tumours.

Cancerous growth as progressive

Carcinogenesis involves a change in the developmental behaviour of a group of cells. Initially, it often involves a loss of some of the functional and structural characteristics that previously appeared in the cells. It is commonly followed, however, by the assumption of new properties, which however untoward they may be for the host animal, must be considered as a progressive type: the cells often grow faster and multiply sooner than the noncancerous cells, for example. Furthermore, the cells may undergo a sequence of changes in character and in their arrangement within the tumour. All these features can be regarded in a developmental sense as progressive.

In view of the great rarity of cases of abnormal development that lead to progressive changes, it seems to follow that the organs produced during the normal development of any given species actually exhaust all the potentialities of its genotype for the production of orderly functional structures. It appears that the only abnormal developments that can be produced are either displacements of normal organs, or inadequacies in carrying out normal processes, or the initiation of progressive but quite disorderly processes, as in the production of tumours (see above *Aberrations of growth: biological malformation*).

GENERAL SYSTEMS OF DEVELOPMENT

Development of single-celled organisms. In viruses, activities consist in the production, aided by the machinery of a host cell, of units for building new virus or phage particles: development is simply the assemblage of these constituent units.

In the next higher grade of biological organization, the organism consists of a single cell. Many single-celled algae produce special forms of cells that correspond to the sex cells, or gametes; these cells may unite in fertilization, the resulting fertilized egg, or zygote, undergoing a short period of development. In many other single-celled organisms, however, reproduction takes place by the simple division of an original cell into two daughter cells. In such forms, development normally is part of the process of subdivision. It involves the remodelling of the parent cell into two smaller cells, which are then separated by the division. Something similar must, of course, be involved in the division of cells of higher organisms also. In many single-celled organisms, however, the cell contains a number of defined parts, which are arranged in very definite ways, so that the process of remodelling is very striking and easily observed. This is so, for instance, with ciliated protozoans, in which the cortex is provided with a large number of hairlike cilia or other appendages, arranged in precise patterns, and often with such other structures as a mouth or a gullet. These structures are reproduced in two identical but smaller copies during cell division. This does not necessarily imply that no other developmental processes are possible. The process of regeneration of parts removed occurs quite independently of cell division, for example.

Protozoan development

Open and closed systems of development. There is a marked difference between the general system of development in multicellular plants and multicellular animals. In a plant, certain groups of cells retain throughout the whole life of the plant an embryonic capability to give rise to many types of cells. These regions, known as meristems, occur at the growing tips of branches and roots and as a cylindrical sheath around the stem. They consist of rapidly dividing cells capable of assembling into groups that form buds from which may arise new stems, leaves, flowers, or roots.

By contrast, most animals have no special regions that retain an embryonic character. In most forms, the whole egg, and the whole collection of cells immediately derived from it, take part in the developmental processes and form parts of the developing embryo. In some forms that go through a number of larval stages, the development of certain cells is interrupted at an early stage, and they are set aside and resume their development to form a later type of larva, or to form the adult after the larval stages are completed. An example would be the imaginal buds of some insects. The cells of these buds cannot be regarded as retaining a fully embryonic character comparable to that of the plant meristems, since they cannot perform all the developmental processes but only those involved in the production of the particular late-larval or adult structure for which they have been set aside. In general, then, plants remain embryonic in character, capable as it were of starting again from the beginning to carry out the entire developmental process. Their development is, in this sense, "open." Most animals, on the other hand, lack persistently embryonic cells of this kind, and their development may be characterized as "closed." (There may be certain exceptions to this in very simple forms, such as flatworms, in which certain cells called neoblasts seem able to participate in any type of development; these cells are usually scattered throughout the body, and the major developmental processes that bring into being the general form of the organism cannot be attributed to them, as the development of the plant can be attributed to the meristems.)

Blastogenesis versus embryogenesis. Some animals possess a second system of development, in contrast to the "closed" embryonic system emphasized in the last section. In its most fully developed form, this system consists in remodelling a portion of the parental body into a new organism without any involvement of eggs or sperm. In an adult hydra, a microscopic aquatic animal, a portion of the body may begin to grow exceptionally fast; its cells differentiate into the various cell types and become molded into the constituent organs to build up a new individual identical to the parent. The group of cells responsible for this behaviour is, in its early stages, referred to as a bud, or blastema. Before they become activated these cells may appear quite indistinguishable from the other cells of the body and betray no embryonic capability comparable to the meristems of plants.

In some higher organisms, including certain insects, reptiles, and amphibians, incomplete but still fairly extensive new developments of a similar kind may take place. They require the stimulus of an injury, however, which may involve the removal of part of the normal body. The usual result is a new development to regenerate, or replace, the missing part. The first stage in such regenerative processes consists in the formation of a blastema, that is, a group of rapidly dividing cells that shows little sign of cellular specialization. The evidence indicates that they may not arise, as was once thought, from persisting embryonic cells scattered within the adult body, but instead are formed of cells near the position of the injury. These cells lose their normal adult character and become capable of developing into most of the tissues required to replace the parts removed by the injury (see above *Biological regeneration*).

The developmental aspect of regeneration

Development from a blastema, or blastogenesis, presents many contrasts to embryogenesis, the normal form of development from a fertilized egg. In blastogenesis, tissues that, during embryonic development, appear in se-

quence one after another, may be formed simultaneously and without any obvious sequential relations. Very little, however, is as yet understood about the mechanisms by which the various tissues within the blastema become differentiated from one another. It may well be that these mechanisms are more similar to those found in embryonic development than appears at first sight.

CONSTITUENT PROCESSES OF DEVELOPMENT

Size and number

Growth. As was pointed out earlier, developing systems normally increase in size, at least during part of their development. "Growth" is a general term used to cover this phenomenon. It comprises two main aspects: (1) increase in cell numbers by cell division and (2) increase in cell size. These two processes may in some examples occur quite separately from each other; for instance, cells in certain rapidly growing tissues (*e.g.*, the connective tissue or blood-forming systems in vertebrates) may increase greatly in number, while the cells remain approximately the same size. Alternatively, in some organs (*e.g.*, the salivary glands of insects) the cells may increase greatly while remaining the same in number, each cell becoming enlarged, or hypertrophied. In such greatly enlarged cells there is often duplication of the genes, involving an increase in the DNA content of the nucleus, although no cell division takes place, and the nucleus continues as a single body, although with a multiplied, or "polyploid," set of chromosomes.

In very many cases, however, the growth of an organ depends on increases both in cell number and in cell size. The relative importance of these two processes has yet to be properly investigated. One case that has been well studied is the size of the wings of the fruit fly *Drosophila*. The number of cells in the wing can be easily determined, since each bears a single hair that can be seen and counted in simple microscopic preparations. It has been found that there is an accommodation of factors: if there is an unusually large number of cells, these may be somewhat smaller than usual, so that the total size of the wing remains relatively unchanged.

Perhaps the major theoretical difficulty in the concept of growth is that it is a quantitative notion attached to an ill-defined entity. Growth is an increase in size; but size of what? If a cell or organ increases in volume merely by the absorption of water, or by the laying down of a mineral substance such as calcium carbonate, is this to be regarded as growth or not?

Morphogenesis. As was pointed out earlier, morphogenesis refers to all those processes by which parts of a developing system come to have a definite shape or to occupy particular relative positions in space. It may be regarded as the architecture of development. Morphogenetic processes involve the movement of parts of the developing system from one place to another in space, and therefore involve the action of physical forces, in contrast to processes of differentiation (see below), which require only chemical operations. Although in practice the physical and chemical processes of development normally proceed in close connection, for purposes of discussion it is often convenient to make an artificial separation between them.

There is an enormous variety of different kinds of structures within living organisms. They occur at all levels of size, from an elephant's trunk to organelles within a cell, visible only with the electron microscope. There is still no satisfactory classification of the great range of processes by which these structures are brought into being. The following paragraphs constitute a tentative categorization that seems appropriate for the present state of biological thought on this topic.

Morphogenesis by differential growth. After their initiation, the various organs and regions of an organism may increase in size at different rates. Such processes of differential growth will change the overall shape of the body in which they occur. Processes of this kind take place very commonly in animals, particularly in the later stages of development. They are of major importance in the morphogenesis of plants, where the overall shape of the plant, the shape of individual leaves, and so on, depends primarily on the rates of growth of such component elements as the stems, the lateral shoots, and the vein and intervein material in leaves. In both animals and plants, such growth processes are greatly influenced by a variety of hormones. It is probable that factors internal to individual cells also always play a role.

Although differential growth may produce striking alterations in the general shape of organisms, these effects should probably be considered as somewhat superficial, since they only modify a basic pattern laid down by other processes. In a plant, for instance, the fundamental pattern is determined by the arrangement of the lateral buds around the central growing stem; whether these buds then grow fast or slowly relative to the stem is a secondary matter, however striking its results may be.

Morphogenetic fields. Many fundamental processes of pattern formation (*e.g.*, the arrangement of lateral buds in growing plants) occur within areas or three-dimensional masses of tissue that show no obvious indications of where the various elements in the pattern will arise until they actually appear. Such masses of tissue, in which a pattern appears, have been spoken of as "fields." This word was originally used in the early years of the 20th century by German authors who suggested an analogy between biological morphogenetic fields and such physical entities as magnetic or electromagnetic fields. The biological field is a description, but not an explanation, of the way in which the developing system behaves. The system develops as though each cell or subunit within it possessed "positional information" that specifies its location within the field and a set of instructions that lays down the developmental behaviour appropriate to each position.

The appearance of pattern

There have been several attempts to account for the nature of the positional information and of the corresponding instructions. The oldest and best known of these is the gradient hypothesis. In many fields there is some region that is in some way "dominant," so that the field appears as though organized around it. It is suggested that this region has a high concentration of some substance or activity, which falls off in a graded way throughout the rest of the field. The main deficiency of the hypothesis is that no one has yet succeeded in identifying satisfactorily the variables distributed in the gradients. Attempts to suppose that they are gradients of metabolic activity have, on investigation, always run into difficulties that can only be solved by defining metabolic activity in terms that reduce the hypothesis to a circular one in which metabolic activity is defined as that which is distributed in the gradient.

Recently, a new suggestion has been advanced concerning position information. Most processes within cells normally involve negative feedback control systems. These systems have a tendency to oscillate, or fluctuate regularly. In fact, any aspect of cell metabolism may be basically oscillatory in character; the cycle of cell growth and division may be only one example of a much more widespread phenomenon. The substances involved in these oscillations are likely to include diffusible molecules capable of influencing the behaviour of nearby cells. It is easy to envisage the possibility that there might be localized regions with oscillations of higher frequency or greater amplitude that act as centres from which trains of waves are radiated in all directions. It has been suggested that positional information is specified in terms of differences in phase between two or more such trains of transmitted oscillations.

Certain types of field phenomenon may involve an amplification of stochastic (random) variations. In systems containing a number of substances, with certain suitable rates of reaction and diffusion, chance variation on either side of an initial condition of equilibrium may become amplified both in amplitude and in the area involved. In this way, the processes may give rise to a pattern of differentiated areas, distributed in arrangements that depend on the boundary conditions.

Morphogenesis by the self-assembly of units. Complex structures may arise from the interaction between units that have characteristics such that they can fit together in a certain way. This is particularly appropriate for morphogenesis at the simple level of molecules or cells. Units such as the atoms of carbon, hydrogen, oxygen, nitrogen, and so on, can assemble themselves into orderly molecular struc-

Molecular assembly

tures, and larger molecules, such as those of tropocollagen, or protein subunits in general, can assemble themselves into complexes whose structure is dependent on localized and directional intermolecular forces. It seems that such comparatively large entities as the units that come together to form the head structures of bacteriophages or bacterial flagella are capable of orderly self-assembly, but the chemical forces that give rise to the interunit bonds are still little understood.

Processes that fall into the same general category as self-assembly may occur within aggregates of cells. The units that self-assemble are the cells themselves. Interaction and aggregation may be allowed to occur in assemblages of cells of one or more different kinds. In such cases it is commonly found that the originally isolated cells tend to adhere to one another, at first more or less at random and independently of their character, but later they become rearranged into a number of regions consisting of cells of a single kind. When the cells in the initial collection differ in two different characteristics, for instance in species and organ of origin, the assortment in some cases brings together cells from the same organ, in other cases cells from the same species. Mixtures of chick and mouse cells, for instance, reassort themselves into groups derived from the same organ, whereas cells from two different species of amphibia sort out into groups from the same species more or less independently of organ type.

This morphogenetic process probably has only a restricted application to the formation of structures in normal development, in which only in a few tissues (*e.g.*, the connective system) do cells ever pass through a free stage in which they are not in intimate contact with other cells, and cells of different origin do not normally become intermingled so as to call for processes of reassortment. To explain normal morphogenetic processes of plants and animals one must look to the results that can be produced by the differential behaviour of cells that remain in constant close contact with one another. Several authors have shown how striking morphogenetic changes could be produced within a mass of cells that remain in contact, but that undergo changes in the intensity of adhesion between neighbouring cells, in the area of surface in the proportion to cell volume, and so on.

Differentiation. Differentiation is simply the process of becoming different. If, in connection with biological development, morphogenesis is set aside as a component for separate consideration, there are two distinct types of differentiation. In the first type, a part of a developing system will change in character as time passes, for instance, a part of the mesoderm, starting as embryonic cells with little internal features, gradually develops striated myofilaments, and with a lapse of time develops into a fully formed muscle fibre. In the second type, space rather than time is involved; for instance, other cells within the same mass of embryonic mesoderm may start to lay down an external matrix around them and eventually develop into cartilage. In development, differentiation in time involves the production of the characteristic features of the adult tissues, and is referred to as histogenesis. Differentiation in space involves an initially similar (homogeneous) mass of tissue becoming separated into different regions and is referred to as regionalization.

Tissues and regions

Histogenesis involves the synthesis of a number of new protein species according to an appropriate timetable. The most easily characterized are those proteins formed in a relatively late stage of histogenesis, such as myosin and actin in muscle cells. The synthesis of proteins is under the control of genes, and the problem of histogenesis essentially reduces to that of the genetic mechanisms that direct protein synthesis.

Regionalization is concerned with the appearance of differences between various parts of what is at first a homogeneous, or nearly homogeneous, mass. It is a prelude to histogenesis, which then proceeds in various directions in the different regions so demarcated. The processes by which the different regions acquire distinct contrasting characteristics must be related to some of the processes discussed under morphogenesis. Unlike morphogenesis, regionalization need not involve any change in the overall spatial shape of the tissues undergoing it. Regionalization falls rather into the type of process for which field theories have been invoked.

Control and integration of development

PHENOMENOLOGICAL ASPECTS

One of the most striking characteristics of all developmental systems is a tendency to produce a normal end result in spite of injuries or abnormalities that may have affected the system in earlier stages. In many cases, perhaps in most, only injuries inflicted during a certain restricted period of development can be fully compensated for. During such periods the system is said to be capable of regulation or the restoration of normality.

Developmental regulation is often discussed in terms of homeostasis, or regulatory mechanisms. Many systems, including biological ones, exhibit a tendency to return to initial equilibrium once they are diverted from it. A developing system is, by definition, always changing in time, moving along some defined time trajectory, from an initial stage, such as a fertilized egg, through various larval stages to adulthood, and finally to senescence. The regulation that occurs in such systems is a regulation not back to an initial stable equilibrium, as in homeostasis, but to some future stretch of the time trajectory. The appropriate word to describe this process is homeorhesis, which means the restoration of a flow.

A second major phenomenological characteristic of development is that the end state attained is not unitary but can be analyzed into a number of different organs and tissues. The overall time trajectory of this system can, therefore, also be analyzed into a number of component trajectories, each leading to one or another of the end products that can be distinguished in the later stages. A major discovery of the early experiments on developing systems was that, in many cases at least, the different time trajectories diverge from one another relatively suddenly during some short period of development, which usually occurs well before any visible signs of divergence can be seen microscopically or by any other available means of analysis. The most dramatic and influential example of this was provided by studies on the development of the amphibian egg at the time of gastrulation, or formation of a hollow ball of cells. At this time the lower hemisphere of the embryo will be pushed inward (invaginated) to develop into the mesoderm and endoderm, and the upper hemisphere will remain on the surface, expanding in area to cover the whole embryo. Approximately one-third of the upper hemisphere will develop into the nervous system and the remainder into the skin. During the period when these morphogenetic movements of invagination and expansion are occurring, a process takes place by which a portion of the upper hemisphere enters a trajectory toward neural tissue and another part enters a trajectory leading to epidermal development. This process of determination of developmental pathways happens relatively quickly, during a period when the cells of the two different regions appear superficially alike. The occurrence of the determination can in fact be demonstrated only experimentally. Before it occurs, any part of the hemisphere can develop either into neural tissure or into skin. After it has happened, each part can develop only into one or the other of these alternatives.

Determination

It is clear that an adequate theory of development has to account not only for the processes by which a developing system moves along its appropriate time trajectory, but also for the nature of the processes by which the trajectories diverge from one another and become fixed or determined in the developing cells.

The determined state can be transmitted through many cell generations. An example of this transmission can be seen in *Drosophila* flies. The imaginal buds of *Drosophila* are small packets of cells that become separated from the main body of the embryo in the early stages of development. They persist throughout larval life and then enter into the differentiation of adult characteristics when stimulated to do so by the hormones secreted at the time of pupation. These pupation hormones disappear from the

body of the adult insect, and imaginal buds transplanted into the body cavity of an adult undergo many cell generations, but they do not show any signs of differentiating into the specific tissues of the corresponding adult organ. After many generations of proliferation, however, the cells can be transplanted back into a larva ready to pupate; they thus submit to the pupation hormones and differentiation occurs. Through many generations of proliferation the cells have retained the determination as to which adult organ they will develop into when the pupation hormones become available.

Attempts to identify the determining agent have not yet been successful. Experiments on amphibian eggs, however, have given rise to one important general conclusion; namely, that the process of determination can take place only during a certain period of development, in which the cells of the upper half of the amphibian egg are poised between the two alternatives of development into neural tissue or into skin. They are said at this time to be "competent" for one or the other of these types of development. While they are in this state, and only while they are in it, a variety of external agents can switch them into one or the other of the possible pathways. Such a situation may be contrasted with one in which the cells were neutral, or featureless, and required then an external agent to transmit to them the quality of becoming nervous tissue or of becoming skin. This would mean that the reacting cells required information or instructions to be added to them from outside. Such a situation is not characteristic of biological development. Both in highly developed organisms such as amphibians and in simpler ones such as bacteria, the external agents act only as a releaser that switches on one or another process for which all of the necessary information is already incorporated in the cells concerned.

ANALYTICAL ASPECTS

The existence of these developmental phenomena was realized in the first third of this century. During this period, biologists had no clear notion of the fundamental concepts needed to explain development. Developmental biologists, or embryologists, attempted to account for their observations by means of ill-defined notions, such as "potencies" or "organ-forming substances," or by referring to cellular properties that are real enough but obviously in themselves complex and essentially secondary in nature, such as cellular adhesiveness, the capacity of cell surfaces to differentially absorb certain substances, and so on. It was only gradually that developmental biologists came to realize the importance of the demonstration by genetics that nearly all the instructions required for the building of a new organism are contained in the genes that come together during fertilization, and that the small additional amount of information, contained primarily in the ovum, is itself a product of genetic instructions provided in the body of the mother in which the ovum is produced. The fundamental problems of the theory of development are, therefore, to understand how these units interact with one another to form more complex mechanisms that bring about the cellular or tissue behaviours of the different types of developing systems.

The genes as repositories of developmental instructions

In the development of the neural system of vertebrates, for example, a great many genes must be active in controlling the synthesis of particular proteins. In the formation of the wing of a *Drosophila*, the activity of some 20 or 30 genes has been definitely demonstrated, and certainly many more are involved. The action of all these genes, however, must be considered to form a network involving many types of feedback and other interactive loops, the overall result of which is a product in which many components are present in precisely defined concentrations; and further, the developmental process leading to this end result must be buffered or stabilized, in the sense that if the process is diverted from its normal course at an early stage, it returns to some later stage of the normal trajectory. The realization that the basic units of development are genes indicates that a stabilized time trajectory involves the action of tens, if not hundreds, of genes. The realization that biological development is fundamentally an expression of the controlled activities of genes has finally resolved one of the old philosophical controversies about the nature of development, between preformation and epigenesis. The former supposed that, at the initiation of development, for instance in the fertilized egg, the system already contained some representative of every organ that would eventually put in an appearance. The vindicated theory of epigenesis, on the other hand, supposed that later appearing entities were produced during the course of development.

The modern interpretation of epigenesis is that the initial stage of development does contain certain entities with well-defined properties, namely the genes. These do not, however, represent directly the later formed organs, which arise by the gradual interaction and progressive unfolding of the properties of groups of genes.

One of the major problems confronting modern developmental biology—namely, the nature of "determination"—requires an understanding of how genes are "primed" to enter into activity when an appropriate stimulus is given. The state of priming presumably has to apply to quite a large number of genes, though perhaps not to all that will be involved in the stabilized, or buffered, time trajectory, since some may be brought into activity by the operation of the earlier active ones. The priming, moreover, has to be able to persist through cell division and be capable of transmission through many generations of cell proliferation. Few concrete suggestions as to the mechanism have yet been made. One is that the primed genes are already producing the ribonucleic acid molecules, called messenger RNA's, which direct protein synthesis in the cell, but that these messengers are in some way inactivated or prevented from activating the protein-synthesizing machinery; this is known as the "masked messenger" hypothesis. Arguments in favour of this hypothesis are, however, circumstantial rather than direct. In some cases, for instance that of the *Drosophila* imaginal buds, there is direct evidence against it. Another hypothesis, perhaps more attractive, but much vaguer, is that the determination or priming involves the intervention of some of the large amounts of reiterated DNA known to be present in the cells of higher organisms. At the present time, however, biology lacks any convincing theory of determination in terms of gene action.

The "masked messenger" hypothesis

It appears at first sight that more is known about actual differentiation than initial determination. Actual differentiation must involve the controlled synthesis of particular proteins, coded for by specific genes. Certainly, a great deal is known about the mechanisms that control the action of genes in directing the synthesis of proteins in simple organisms such as viruses and bacteria. It is tempting to suppose that similar systems operate in controlling the synthetic activities of genes in higher organisms. Unfortunately, no single case of an exactly similar controlling system has ever been discovered in higher organisms, in spite of an intense search for it. It may in fact be suggested that until there is a fuller understanding of the mechanism of "priming" genes at the time of determination, there can scarcely be an adequate account of the way in which the activity of these genes is controlled at later stages.

Development and evolution

Evolution is carried out by a process dependent on mutation and natural selection. Expositions of this thesis, however, tend to overlook the fact that mutation occurs in the genotype, whereas natural selection acts only on the phenotype, the organism produced. It follows from this that the theory of evolution requires as one of its essential parts a consideration of the developmental or epigenetic processes by which the genotype becomes translated into the phenotype. The consequences of such considerations are discussed in the following sections.

EFFECT ON LIFE HISTORIES

Length and timing of the reproductive phase. Natural selection results in the production by one generation of offspring that are able to survive and reproduce themselves to form a further generation. The time unit appropriate to natural selection is therefore the generation interval. There will always be some natural selective pressure for the shortening of the generation interval, simply out of

a natural economy, and for an increase of the number of offspring produced by any reproducing individual. One of the ways in which such an increase could be assured would be the lengthening of the reproductive phase in the life history; another would be an increase in the number of offspring produced.

These are, of course, not the only natural selective pressures that operate. It is clear enough that, in evolution, they have often been overcome by other pressures. There is another natural selective pressure of more general importance. This is the pressure to restrict the length of the reproductive period, and indeed to remove reproductive individuals, in order to make room for the maturation of a new generation in which new genetic combinations can be tried out for their fitness. A species whose individuals were immortal would exhaust its possibilities for future evolution as soon as its numbers saturated all the ecological niches suitable for its way of life. Death is a necessary condition for the trying out of new genetic combinations in later generations. It is usually brought about, in great part at least, by combinations of two processes: restriction of the period of effective reproduction to a certain portion of a life history, and as a necessary consequence of this, the absence of natural selection for genetic mutations that would be effective in preserving life after reproduction has ceased. In some organisms—for instance, long-lived trees—there may be no restriction of reproduction to a particular period in the life history, but their development involves the gradual accumulation of larger and larger quantities of nonliving materials, such as dead wood, which presents a growing handicap, in the face of which the organism cannot indefinitely maintain itself against the inevitable hazards of existence. It is still something of a question whether these natural selective forces are sufficient in themselves to account for the phenomena of senescence, aging, and eventual death, which are found in various forms throughout nearly the whole biological kingdom (see below *Aging and senescence*).

The evolutionary significance of death

As was mentioned above, evolution has produced a number of the types of multiphasic development, in which the life history involves a succession of larval stages. Such types of development offer the possibility of changing the relative importance of the various stages in relation to the exploitation of resources and reproduction by the species. There are, for instance, many types of animals (particularly insects) in which nearly the whole life history is passed in a larval stage in which most of the feeding and growth of the organism is carried out, the final adult stage being short and used almost entirely for reproduction. Another evolutionary strategy has been to transfer the reproductive phase from the final stage of the life history to some earlier larval stage. This again has occurred in certain insects. If such a process is carried to its logical evolutionary conclusion, the final previously adult stage of the life history may totally disappear, the larval stage of the earlier evolutionary form becoming the adult stage of the later derivative of it. An example in which this process is at least partially accomplished is in the axolotl, a salamander that reproduces in a larval stage and in nature rarely if ever metamorphoses into the adult, but can be persuaded to do so if injected with extra supplies of the hormone thyroxin. It has been suggested that such processes of neoteny (the retention of some juvenile characteristics in adulthood) have played a decisive role in certain earlier phases of evolution, evidence of which is now lost. It has been argued that the whole vertebrate phylum may have originated from modifications of one of the larval stages of an invertebrate group.

The role of neoteny in evolution

Recapitulation of ancestral stages. The modifications of life histories just mentioned are aspects of a more general situation; namely, that the only variations that can become available for natural selection to operate on are those that can be produced by alterations of the developmental or epigenetic system of an existing organism. Any new mutant gene can cause a change only in a preexisting set of developmental interactions; the phenotypes to which it can give rise are limited by the nature of the system that it will modify. One immediate result of this situation is that the development of a later evolved form will retain many features from the development of its ancestors: most evolutionary developments are likely to be additions to the previous organization. Since there is evolutionary pressure to reduce the length of time between generations, the addition of a new feature to development is likely to be accompanied by a speeding up of the older stages, and probably omission of certain of them.

To repeat, the development of a late-evolved form retains those aspects of earlier life histories that are essential for the building up of later developmental stages that may be important for natural selection. In the vertebrates, for instance, highly evolved types such as mammals and birds produce during their early development remnants of the primitive kidneys (pronephros and mesonephros) that functioned as excretory organs in their evolutionary ancestors. Although these organs no longer perform their physiological functions in later organisms, they play an essential role during the formative processes of embryonic development. Some structures characteristic of evolutionary ancestors may be retained for relatively short evolutionary periods after they have lost their original function simply because there is not sufficient natural selective pressure to bring about their elimination when they no longer have any obvious function, either physiologically or epigenetically; the human appendix is an example.

Adaptability and the canalization of development. A developing organism is subjected to natural selection by its particular environment. The environment is not the same for all individuals of a population, nor does it necessarily remain the same throughout evolutionary periods of time. An organism can be regarded as having to meet environmental changes that are unpredictable. There are basically two different types of strategy employed, in various proportions in different organisms, to meet this situation. One, perhaps the more obvious, is to evolve a high capacity for modification by environmental circumstances in ways that increase fitness in the environment in question; this is the strategy of increasing adaptability. It is probably true to say that all organisms show some capacity for adaptation, either short-term (physiological) or longer term (developmental), to their environments. In most organisms, however, particularly in most higher organisms, there is considerable development of the alternative strategy, which is to build up well-buffered or channelled developmental processes, which lead to the production of a relatively predictable invariant end result in the face of very diverse environments. The second strategy is likely to be followed in situations in which the environment is likely to change markedly during the course of the organism's life.

The tendency to channel development

Whether or not this is the main reason for the evolution of channelled, or canalized, developmental systems, a considerable degree of canalization is very common. It is relatively rare to find instances in which the form of an animal is highly dependent on the early environment, although such dependence is common enough among plants. Much more frequently, situations such as that typified by the house mouse are encountered: the mouse develops into an almost identical form whether it lives in the tropics or in a cold-storage depot.

This canalization of development severely restricts the phenotypic effects that can be produced by mutations. In particular, many new mutations occurring in a single dose in a diploid organism are found to be recessive, or ineffective in causing any alteration in the phenotype. As this discussion makes clear, canalization should not be considered as a relation involving only the normal and mutated forms of a particular gene, but rather the result of the interaction of many genes.

GENETIC ASSIMILATION

A long-standing controversy in biology has been concerned with whether phenotypic modifications produced by abnormal environments are heritable in the sense that they can be produced by later generations in the absence of the original environmental stress. The hypothesis that they are heritable was advanced by the French evolutionist Lamarck in the 18th century and is generally known as the "inheritance of acquired characters." It found some

supporters among biologists, some of whom used it as an argument against the Darwinian theory of evolution. In a broad sense, all characters are to some extent inherited, in that they depend on the genotype of the organism, and to some extent acquired, since development is also affected by the environment. In a stricter sense, however, Lamarck's hypothesis suggests that there is some inherent biological property that enables organisms to pass on physical modifications to their descendants, independently of a Darwinian mechanism of selection.

The combination of adaptability and canalization in development can explain such phenomena in strictly Darwinian rather than Lamarckian terms. The abnormal environment acting during development may succeed in modifying even a well-canalized development system. If the modification is of an adaptive kind and increases the fitness of the individuals in the unusual environment, it will be favoured by natural selection. The development of the selected individuals will, however, also show some properties of canalization, that is to say, resistance to further environmental changes. This invariance may be sufficient to prevent offspring of the selected individuals from reverting completely to the original phenotype even if they are removed from the abnormal environment. After selection for an adaptive modification in an abnormal environment has proceeded for many generations, a form may be produced whose canalization is strong enough to maintain the new phenotype almost unaltered when the environment reverts to what it was before the abnormality occurred. This process, which has been demonstrated in a number of laboratory experiments, is known as genetic assimilation. It produces exactly the same results as those emphasized by advocates of the Lamarckian inheritance of acquired characters, but it produces them by an orthodox Darwinian mechanism operating on developmental systems that have the common properties of canalization and adaptability. It provides the most convincing explanation for the evolution of organisms that are physiologically or functionally adapted to the demands their way of life will make.

(C.H.W.)

PLANT DEVELOPMENT

Although both plants and animals share the chemical basis of inheritance and of translation of the genetic code into structural units called proteins, plant development differs from that of animals in several important ways. Higher plants sustain growth throughout life and, in this sense, are perpetually embryonic; animals, on the other hand, generally have a determinate period of growth, after which they are considered mature. Furthermore, both growth and organ formation in plants are influenced by their possession of a rigid cell wall and a fluid-filled space called the vacuole, two features unique to the plant cell. Conversely, certain features of animal cells are absent in plants. Notable is the lack of cellular movements and fusions that play an important part in tissue and organ development in higher animals.

General features

LIFE CYCLES

Development of gametes

The life cycle of all tracheophytes (vascular plants), bryophytes (mosses and liverworts), and many algae and fungi is based on an alternation of generations, or different life phases: the gametophyte, which produces gametes, or sex cells, alternating with the sporophyte, which produces spores. Gametophytes develop from the spores and, like them, are normally haploid; *i.e.*, each cell has one set of chromosomes. Sporophytes develop from a fertilized egg, or zygote, that results from the fusion of gametes (fertilization) formed by the gametophytes; they are accordingly diploid; *i.e.*, each cell has two sets of chromosomes. Although the two generations are phases of one life cycle, they have independent developmental histories: each begins as a single cell, passes through a juvenile period, matures, and gives rise to the alternate phase.

In various algae and fungi the two generations are alike in form (*i.e.*, are isomorphic), and, despite the difference in chromosome number, their development follows essentially identical pathways. More commonly, however, the alternating generations have different forms (*i.e.*, are heteromorphic); this is true for the bryophytes and for all vascular plants, both angiosperms (flowering plants) and gymnosperms (conifers and allies). General rules for vascular plants are that the sporophyte generation is physically the larger, has a more complex developmental history, produces a greater range of cell types, and expresses a more diverse biochemistry; the gametophyte is often diminutive, reduced in the case of the angiosperms to a mere few cells. In the bryophytes, the gametophyte generation, rather than the sporophyte, is the more conspicuous.

Although the gametophyte generation in vascular plants is small and has limited physiological capabilities, its cells must convey genes capable of directing the sporophytic developmental pattern, because the pattern is transmitted through the gametes to the zygote (Figure 6). The expression of "sporophytic" genes must therefore be repressed in the gametophyte, probably from the time of spore formation (sporogenesis). Correspondingly, events associated with gamete formation (gametogenesis) or fertilization must somehow free the sporophytic genes and thus permit the zygote to enter the sporophytic developmental pattern. Although it might be supposed that the "switch" is associated with the difference in chromosome number between the haploid spore (a single set) and the diploid zygote (a double set), this has been shown not to be the determining factor.

Conveying the genes through alternate generations

The alternation of generations illustrates an important principle, namely that cell lineages arising from single parental cells containing the same genetic potentiality may pursue mutually exclusive developmental patterns. Channelling, or canalizing, events of this nature occur repeatedly in the course of development of an individual plant, beginning with the pattern of cell division from the very first cleavage of the zygote cell.

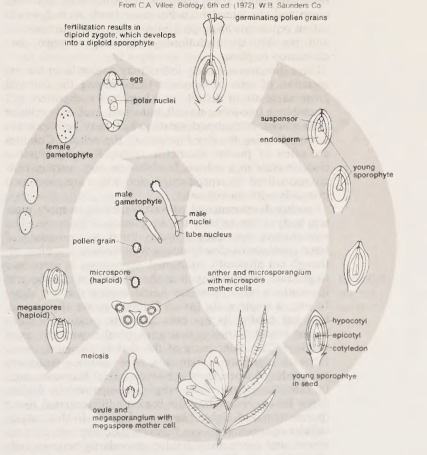


Figure 6: Stages in the fertilization and seed formation of a dicot.

BODY PLANS

Collectively plants manifest a wide range of body plans, ranging from the single cell (or unicell), with a single nucleus, through various types of colonial and filamentous forms to massive multicellular structures. (Algae, including the single-celled forms, have a great deal in common in structure and biochemistry with vascular plants. Bacteria and fungi generally are not considered plants, but because of their various plantlike qualities they are taken as plants for the purposes of this section.)

For the unicell, development is the same as cell differentiation. Although many unicellular fungi and algae show little differentiation other than that connected with reproduction, others undergo elaborate structural changes that illustrate many principles basic to development in multicellular plants. An important example is the green alga *Acetabularia*. This alga first produces a rootlike system and stalk and then, later, a flattened umbrella-like cap. The developmental potentialities of this unicell, with its single nucleus, are, however, limited; in order for there to be any advance beyond the state seen in *Acetabularia*, with the development of greater body mass and a division of labour among different parts, an increase in the number of participating nuclei seems obligatory.

One method of providing more nuclei is by nuclear division without a corresponding cell division; the result is a coenocytic structure. Plants with this type of multinucleate organization show considerable diversity; examples are found in both algae and fungi. Growth occurs by the extension of the cell wall in certain zones, usually at the tips of filaments, and structural differentiation results from branching and the specialization of parts for particular functions. The aggregation of coenocytic filaments can lead to the development of a three-dimensional body, or thallus, but plants with this type of organization have not achieved great size.

A more significant type of body plan, one based on the multicellular filament, is found in its most simple form in certain algae known as diatoms, in which chains of cells of indefinite length arise, although the cells show no evidence of interaction. More advanced is the condition of many other algae, in which there occur branches that may either be identical with the original filament or show structural or physiological specialization. This condition occurs in certain green algae, in which the main branches creep and the lateral branches grow erect; such diversification represents an important developmental innovation and, possibly, the evolutionary beginnings of organ specialization in plants.

Multicellularity: an advance in development

Three-dimensional body forms may evolve from the association of cells in colonies. Cells among the colonial green algae are of definite number; each component cell resembles a free-living unicell, but all are united by cellular connections, or plasmodesmata, which may be important in coordinating the development of the colony. Colonies are often of precise geometric shape, forming either a circular plate or a sphere. In elaborate ones, certain cells are specialized for reproduction, and others are concerned primarily with movement.

Another developmental pathway resulting in more massive body structure is by the association of filaments. The reproductive structures of many fungi are composed of large numbers of closely interwoven filaments, which, although not physically connected, do interact in some way to produce structures such as the mushroom cap. Several filamentous body plans are found among the red algae. In one, a single main (axial) filament grows at one end, but, just behind its tip, cells divide to produce a number of lateral filaments that grow parallel with the axial filament. The older parts of the thallus, therefore, seem to be an aggregate of filaments. More massive structures are produced when there are several axial filaments; and, by branching, particularly when accompanied by fusion, dense tissues resembling the basic undifferentiated tissue (parenchyma) of higher plants are formed. In these algae, cellular connections occur between daughter cells of a filament, and others may develop secondarily between cells of neighbouring filaments.

The transition from a filamentous to a three-dimensional form appears most notably in the brown algae. In certain brown algae, growth is by an axial filament, but, behind the tipmost cell, divisions produce a denser tissue lacking evidence of filamentous organization. In the sporophytes of kelp, one of the largest and most complex of the algae, cell division often is restricted to areas comparable to the growing tips of vascular plants, and, although a filamentous organization may be evident in the centre of the thallus, the surrounding cortical regions are composed of a tissue that is essentially undifferentiated. (The gametophytes of kelp, however, have a simple filamentous organization.)

Three-dimensional body form

Among nonvascular plants, true parenchyma is found in the bryophytes, in both the gametophyte and sporophyte phases. The development of the moss gametophyte illustrates the transition from a filamentous to a highly organized three-dimensional growth form. The moss spore germinates into a filamentous plant, the protonema, which later produces a leafy shoot. This type of transition from simple to more complex growth form is accompanied by the synthesis of new kinds of ribonucleic acids (RNA's), presumably through the activation of genes that were not expressed during the early growth of the gametophyte.

Much of the remainder of this section is concerned with the development of the complex body forms of vascular-plant sporophytes, which do not normally pass through any filamentous stages. It may be noted, however, that, in the course of evolution, the capacity for this type of growth has not been lost, since it may be adopted by cells grown in tissue cultures in the laboratory.

PREPARATORY EVENTS

The sporophytes of all vascular plants produce cells called spore mother cells—since they will give rise to spores—in spore cases (sporangia). Spore mother cells are usually surrounded, during development, by a special nutritive tissue. In the more primitive groups each sporangium holds many mother cells. This is true also in the pollen-producing sporangia of gymnosperms and angiosperms but not in the egg-producing sporangia (ovules), which usually have only one mother cell.

In certain lower vascular plants, typified by the club moss *Selaginella*, the gametophyte is formed entirely—or almost entirely—within the spore wall. Two kinds of gametophytes develop from the two kinds of spores produced by the sporophyte in different sporangia; the larger spore (megaspore) gives rise to the female gametophyte, the smaller spore (microspore) to the male. This condition is referred to as heterospory. The gametophytes, or prothalli, of other club mosses and most horsetails and ferns are sexually undifferentiated and arise from one kind of spore, a condition termed homospory.

In these groups the gametophytes develop as free-living and independent plants that ultimately produce the gametes. In general, the male gametes (antherozoids) are produced in globose structures (antheridia) that are either stalked or sunken in the gametophyte. The antherozoids, always many in number, develop from mother cells enclosed in the jacket of the antheridium. Each antherozoid can move by using its whiplike hairs, or flagella, two or three (in the lycopods) or many (in the horsetails and ferns). The female gametes are formed singly in flask-shaped structures (archegonia) that also are either stalked or sunken in the gametophyte. The neck of the flask is closed by neck canal cells, which later break down to permit the entry of the male gamete. The egg itself lies in the basal part, or venter, of the flask, with a ventral canal cell above it. When the male gametes, or antherozoids, are released by the rupture of the antheridium, they swim in a water film to the archegonia and effect fertilization.

Among the gymnosperms the male gametophyte is much reduced and is a parasite on the sporophyte for only a short time. Cell cleavages within the spore wall cut off a prothallial cell, which will give rise to the vegetative (*i.e.*, nonreproductive) part of the plant, and an antheridial cell, which divides into a tube cell and a generative cell. The male gametophyte so formed and contained within the spore wall is the pollen grain. After transfer to the ovule by wind, the pollen grain germinates to form a tube, and

the generative cell divides into two cells, one of which forms the male gametes by further division. The gametes bear numerous spirally arranged flagella. The female gametophyte meanwhile develops entirely within the parent sporangium in the ovule. The size of the single functional spore increases greatly as the spore nucleus divides repeatedly to produce numerous free nuclei. Cell-wall formation then begins at the periphery, extending inward until the whole area is divided into cells. Up to four archegonia are formed, sunken in the tissue of the gametophyte, each with a female gamete, or egg.

Fertilization: a change of generations

The end of the gametophyte phase and the beginning of the sporophyte phase occur at fertilization, when one of the male gametes fuses with the female gamete to form the zygote, which will then develop as the sporophyte. (Development of the sporophyte can, in some cases, be triggered by means other than fertilization, in which case the organism is said to arise parthenogenetically.)

The male gametophyte of angiosperms is reduced to three cells, one so-called vegetative cell and two male gametes. The division producing the gametes may occur either before dispersal of the pollen grain or later, during the growth of the pollen tube. The female sporangium has one or two coats, or integuments, except for an opening (micropyle) at one end; the sporangium with an integument is called the ovule. The female gametophyte, known in this group as the embryo sac, develops from the parent spore while it is still retained in the sporangium. Three cell divisions result in eight nuclei, which arrange themselves so that three lie at each end and two lie in the centre. The cytoplasm then cleaves and three cells are formed at each pole, leaving two nuclei in a large central cell. The three cells at the micropylar pole (end toward the micropyle) form the egg apparatus. Two of these cells, called synergids, correspond to the neck cells of an archegonium; the third is the egg cell. The three cells at the opposite pole, the antipodals, play a part in embryo nutrition in certain genera. The two polar nuclei in the central cell ultimately unite, becoming the fusion nucleus. The pollen grain is transferred by various agencies (wind, water, animals) to the stigma of the female flower, and, as in the gymnosperms, it germinates to produce a tube. This tube grows through intervening tissues, through an opening (micropyle) of the egg, and enters a cell near the micropyle (synergid), in which the two male gametes are discharged. The unique feature of this phase of angiosperm development is that two fertilizations occur. One male gamete fuses with the egg to give the diploid zygote; the other makes its way to the fusion nucleus in the central cell, already diploid, and by a second fusion gives a triploid primary endosperm nucleus, which is later concerned in the formation of the nutritive tissue, or endosperm.

Early development: from zygote to seedling

EMBRYO FORMATION

Cleavage of the zygote. In vascular plants embryo formation, or embryogenesis, usually occurs within a few hours after fertilization, with the first cell division that cleaves the zygote, or fertilized egg, into two daughter cells. Thereafter, rapid cell division provides the building blocks of the primary organs of the embryo sporophyte: the first root, first leaves, and the shoot apex. Temporary structures concerned with embryo nutrition—suspensor and foot—may also be produced. These organs originate in a polarization established at the time of zygote cleavage, but the details of their development vary widely among the different groups.

In the club mosses the zygote divides in a plane at right angles to the axis of the archegonium. The daughter cell toward the neck forms a short filament of cells, the suspensor; the inner cell gives rise to the other organs of the embryo, the shoot, root, and foot. The axis of the embryo is inclined to that of the archegonium and may be almost at right angles. This is in contrast to the behaviour of the true mosses, in which the embryo is oriented along the length of the archegonium, with the foot directed inward and the structures that are equivalent to the shoot, namely the spore capsule and its stalk, directed toward the neck.

A polarity like that of the mosses appears in the horsetails, in which the zygote divides by transverse and longitudinal walls to form a group of four cells. Of these, the two cells toward the neck give rise to the shoot system; the inner two produce the foot and root.

Embryogenesis in gymnosperms and angiosperms

The details of early embryogenesis in gymnosperms vary consider-ably. In the cycads and ginkgos, the initial cleavage establishes a polarity opposite to that in the horsetails, the inner cell giving rise to the shoot and the outer producing the root. Many conifers are unique in that the zygote undergoes a period of free-nuclear division without cell formation, producing usually four or eight nuclei, which move to the end of the zygote, away from the neck cells, where cleavage begins. In the pines a further division gives four tiers of four cells. The intermediate tiers extend greatly to form a suspensor; each of the four cells at the lower pole may act as the parent cell of an embryo, a condition sometimes referred to as polyembryony.

In contrast, there is no free-nuclear stage in angiosperm embryogenesis. The zygote cleaves by a wall more or less at right angles to the axis of the embryo sac. The daughter cell next to the micropyle (basal cell) produces a suspensor and contributes to the root; the inner (terminal) cell gives rise to the shoot system. (Angiosperm embryogenesis is more fully described in the following section dealing with the origin of primary organs.)

Notwithstanding the variation in the different groups, the pattern of development established in the early cell cleavages is consistent. The primary polarization of the zygote must necessarily be imposed by the adjacent tissues of the sporophyte, but thereafter the fate of daughter cells depends on control established within the young sporophyte itself.

Although it is often possible to specify the origin of the cell lineages contributing to the various organs and tissue layers, a geometric regularity in cell division is generally maintained through only the first few division cycles in the embryo. The final form of the embryo is thus determined not through the specification of a precise scheme of cell division, as in the development of colonial algae, but through an overall control in which cell and tissue interactions play an important part.

Origin of the primary organs. Angiosperm embryogenesis can be described in terms of a much studied flowering plant called shepherd's purse (*Capsella bursa-pastoris*). The zygote divides into two cells, the terminal cell and the basal cell. The terminal cell divides by a wall formed at right angles to the first cleavage wall and then again by a wall formed at right angles to this; a quadrant of cells is thus formed (Figure 7A–D). The partition of the quadrant cells in a transverse plane then produces an octant stage (Figure 7E). By transverse divisions, the basal cell forms a filament, the suspensor, of up to ten cells, the end cell of which swells to form an absorbing organ. The attachment cell, or hypophysis, adjoins the octants derived from the terminal cell (Figure 7F).

At this time, the prospective future of each of the zones of the embryo can be specified. Four cells of the octant group will ultimately produce the seed leaves (cotyledons) and the shoot apex; the other four will form the hypocotyl, the part of the embryo between the cotyledons and the primary root (radicle). The hypophysis will give rise to the radicle and the root cap; the cells of the suspensor will degenerate as the embryo matures.

Establishing the embryo

The zones of the embryo destined to form the principal organs are established by this first sequence of divisions, and tissue layers are defined during the ensuing divisions. The octant cells divide by curved walls parallel to the surface; in this way the outer layer responsible for producing the epidermis of the shoot system is defined. Divisions of a more irregular nature in the inner zone ultimately define the tissues from which the central cylinder and vascular core of the main axis of the shoot will develop. Simultaneously, the hypophysis forms a group of eight cells by three successive divisions, the planes of which are mutually at right angles. Of these eight cells, the outer four produce the root cap and epidermis; the inner four contribute to the radicle.

The embryo is at first globular (Figure 7F), but it soon

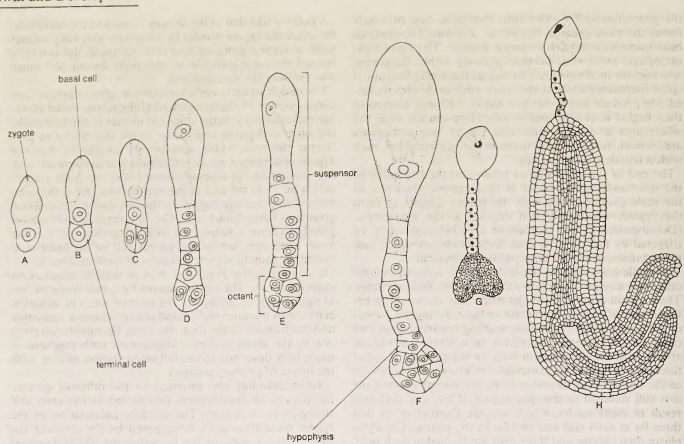


Figure 7: Development of the embryo in shepherd's purse *Capsella bursa-pastoris* (see text; G and H less enlarged than A–F).

After Soueges and Schaffner in P. Maheshwari, *Introduction to the Embryology of the Angiosperm*, copyright © 1950; used with permission of McGraw-Hill Book Company

becomes heart-shaped by a combination of numerous cell divisions and enlargement in two zones of the outer hemisphere (Figure 7*G*). In this manner two cotyledons form. The volume of tissue between the cotyledons is the prospective shoot apex. The characteristic form of the apex is not established until after germination.

As the cotyledons become extended, the embryo bends, because of physical restraints, to conform with the cavity of the embryo sac (Figure 7*H*). From the heart-shaped phase onward, the core of the hypocotyl and the radicle appears as a cylinder of narrow and elongated cells. This is the parent tissue of the vascular system of the seedling. The surrounding tissue contributes the cortex layer of the stem and root.

The embryogenesis of *Capsella* illustrates only one of several patterns found among flowering plants. Among dicotyledons, the planes of division of the terminal cell, the form of the suspensor, and the contribution made by the basal cell to the embryo all provide evidence used in determining the embryogenetic plan.

Monocotyledons, flowering plants the seeds of which contain only one cotyledon, share with dicotyledons such as *Capsella* the main features of early embryogenesis, including the possession of a suspensor and, in most cases, a fairly regular progression of cell divisions to the octant stage. Thereafter the symmetrical growth pattern is lost through the development of the single cotyledon. In the lily family (Liliaceae), generally accepted as a primitive family of monocotyledons, the cotyledon is derived from an octad of cells arising from the terminal cell. The hypocotyl and stem apex are derived from the proximal cell of a short filament formed by the basal cell, and the root comes from the pair of cells next to it. The suspensor forms from the distal cell or cells of the filament. In the more advanced families of monocotyledons, including the grasses (Gramineae) and orchids (Orchidaceae), embryogenesis is much less regular. The grass embryo possesses structures that do not occur in any other flowering plants, namely, the scutellum, an organ concerned with the nutrition of the seedling, and the coleoptile and coleorhiza, protective sheaths of the young shoot and the radicle. The scutellum arises from octant cells, which also contribute to the cotyledon. The basal cell forms part of the coleoptile and also gives rise to the shoot apex and the tissues of the root and coleorhiza. The embryo is asymmetrical, with the shoot apex lying on one side in a notch, ensheathed by the coleoptile.

In marked contrast, embryogenesis of the orchids is more simple. Except when a suspensor is formed, early cleavages follow no well-defined plan, and the product is an ovoid mass of tissue called the proembryo. No cotyledon, stem apex, or root apex is organized in this early period; these organs do not appear until after germination has occurred.

Nutritional dependence of the embryo. During their early growth, the embryos of all vascular plants exist as virtual parasites depending for nutrition on either the gametophyte or the previous sporophyte generation through the agency of the gametophyte or, in the special case of the angiosperms, upon an initially triploid tissue, the endosperm, which is itself nourished by the parent sporophyte.

Embryos as parasites

The early nutrition of the sporophyte in ferns, horsetails, and club mosses such as *Lycopodium* is clearly provided by the gametophyte. In these groups the young sporophyte produces a multicellular structure, the foot, which remains embedded in the tissues of the gametophyte throughout early development withdrawing nutrients. Ultimately, both shoot and root of the sporophyte grow out from the gametophyte, but, even after the first leaf has begun to photosynthesize and thus to produce its own food, the gametophyte may persist.

In *Selaginella*, the gametophytes are sexually distinct. The female gametophyte develops within the wall of the megaspore. The archegonia are exposed after the megaspore wall splits, but the gametophyte never escapes completely. After fertilization, the zygote cleaves, and the outer cell produces a long suspensor that pushes the embryo deeply into the tissues of the gametophyte. A foot is then formed, as in *Lycopodium*, and further development of the embryo continues at the expense of reserves transferred to the megaspore from the preceding sporophyte generation.

There are superficial similarities between the nutritional history of the embryo in gymnosperms and in *Selaginella*, for, in each, the female gametophyte, dependent upon reserves derived from the sporophyte, acts as an intermediary between one sporophyte generation and the next.

In the pines, the female gametophyte develops within the tissues of the nucellus and acquires abundant food reserves. The proembryo forms after a period of free-nuclear division in the zygote, and the tier of cells above the basal

four then elongates to form a suspensor, which pushes the embryonic group deep into the gametophyte (Figure 8). Secondary suspensor cells may form from the basal tier to continue the process. During embryogenesis, the gametophyte continues to grow and to accumulate food materials, which are transferred to the embryo or remain as reserves in the seed.

From *Comparative Morphology of Vascular Plants* by Adriance S. Foster and Ernest M. Gifford, Jr. W.H. Freeman and Company. Copyright © 1959

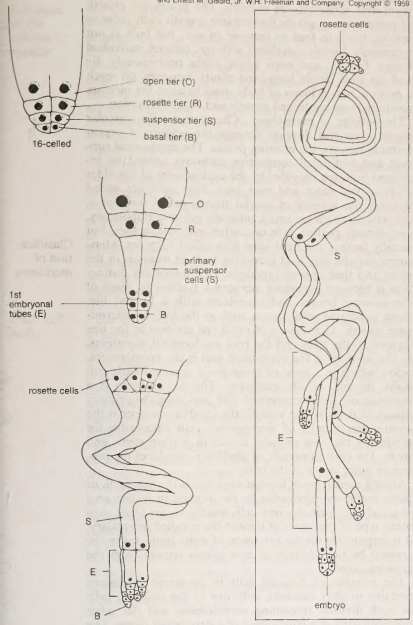


Figure 8: Fertilization and early embryogenesis in *Pinus*.

The female gametophyte of angiosperms never acquires copious reserves, although starch is frequently present in the central cell and sometimes in the egg itself. The unique feature, here, is that the embryo is nutritionally dependent upon the endosperm, a tissue that, in the genetical sense, constitutes a third organism—neither gametophyte nor sporophyte. Furthermore, as a tissue the endosperm manifests several other special characteristics. The nuclei have three chromosome sets and, therefore, three times the deoxyribonucleic acid (DNA) of haploid cells. As nuclear division ends, the amount of DNA per nucleus increases still further, a condition comparable with that in various plant- and animal-gland nuclei, presumably connected with the nutritional function of the endosperm. Nuclear division takes place at first without cell-wall formation so that a coenocyte is produced; later, partitioning of the cytoplasm results in a cellular tissue.

The food reserves

The reserves accumulated in the endosperm include carbohydrates (especially starch), lipids, and proteins. As reserves accumulate, the nuclei of the endosperm cells may undergo deformation and degeneration. In many plants the growing embryo consumes the endosperm before seed maturation; in others, the tissue persists in the seed, providing a reserve for the developing seedling after germination. Endosperm is not formed in certain angiosperms. In such cases the embryo depends on the transfer of nutrients directly from the sporophyte.

Tissues other than the endosperm may become specialized for the early nutrition of the embryo. The antipodal cells of the female gametophyte sometimes acquire glandular properties, as may cells of the nucellus surrounding the embryo sac. In some species the embryo itself develops a suspensor that penetrates the tissues of the parent sporophyte and acts as an absorbing organ.

Dormancy of the embryo. Among the lower pteropsids (club mosses, horsetails, and ferns), the principal agent of dispersal is the haploid spore and not, as in gymnosperms and angiosperms, the seed, the ripened ovule containing a dormant embryo. Since the embryo of lower pteropsids is not involved in dispersal, it does not usually undergo any marked period of dormancy after the differentiation of the primary organs. Development instead proceeds continuously through dependence upon the gametophyte until the young sporophyte is established as a physiologically independent plant. The embryos of gymnosperms and angiosperms pass into a state of dormancy soon after the differentiation of the primary organs and the sporophyte is dispersed in a seed.

In the period leading up to dormancy, several changes occur in the embryo. The accumulation of reserves in the cotyledons or elsewhere ceases, respiratory rate declines rapidly, and cell division, with associated protein and nucleic-acid synthesis, stops. Correlated with these events are cellular changes typical of tissues with low metabolic activity. Especially obvious is the general dehydration of the cells that constitute the seed and the thickening of the cell walls of the ovule to form the seed coat (testa). The product is a structure in which the embryo is protected from temperature extremes by its state of desiccation and is often guarded from further drying and from mechanical or biological degradation by the seed coats. The seed coat often contributes to the maintenance of dormancy by physically impeding the passage of water and gases to and from the embryo, by chemically inhibiting germination, and by mechanically restricting the growth of the embryo.

Role of the seed coat in dormancy

GERMINATION AND EARLY GROWTH

Dormancy is brief for some seeds, for example those of certain short-lived annual plants. After dispersal and under appropriate environmental conditions, such as suitable temperature and access to water and oxygen, the seed germinates, and the embryo resumes growth.

The "breaking" of dormancy. The seeds of many species do not germinate immediately after exposure to conditions generally favourable for plant growth but require a "breaking" of dormancy, which may be associated with change in the seed coats or with the state of the embryo itself. Commonly the embryo has no innate dormancy and will develop after the seed coat is removed or sufficiently damaged to allow water to enter. Germination in such cases depends upon rotting or abrasion of the seed coat in the soil. Inhibitors of germination must be either leached away by water or the tissues containing them destroyed before germination can occur. Mechanical restriction of the growth of the embryo is common only in species that have thick, tough seed coats. Germination then depends upon weakening of the coat by abrasion or decomposition.

In many seeds the embryo cannot germinate even under suitable conditions until a certain period of time has lapsed. The time may be required for continued embryonic development in the seed or for some necessary finishing process—"after ripening"—the nature of which remains obscure.

Low temperatures and seed germination

The seeds of many plants that endure cold winters will not germinate unless they experience a period of low temperature, usually somewhat above freezing. Otherwise germination fails or is much delayed, with the early growth of the seedling often abnormal. (This response of seeds to chilling has a parallel in the temperature control of dormancy in buds.) In some species, germination is promoted by exposure to light of appropriate wavelengths; in others, light inhibits germination. For the seeds of certain plants, germination is promoted by red light and inhibited by light of longer wavelength, in the "far red" range of the

spectrum. The precise significance of this response is as yet unknown, but it may be a means of adjusting germination time to the season of the year, or of detecting the depth of the seed in the soil. Light sensitivity and temperature requirements often interact, the light requirement being entirely lost at certain temperatures.

In the process of germination, water is absorbed by the embryo, which results in the rehydration and expansion of the cells. Shortly after the beginning of water uptake, or imbibition, the rate of respiration increases, and various metabolic processes, suspended or much reduced during dormancy, resume. These events are associated with structural changes in the organelles (membranous bodies concerned with metabolism), in the cells of the embryo.

The emergence of the seedling. Active growth in the embryo, other than swelling resulting from imbibition, usually begins with the emergence of the primary root from the seed, although in some species (*e.g.*, the coconut) the shoot emerges first. Early growth is dependent mainly upon cell expansion, but, within a short time, cell division begins in the radicle and young shoot; thereafter, growth and further organ formation (organogenesis) are based upon the usual combination of increase in cell number and enlargement of individual cells.

Until it becomes nutritionally self-supporting, the seedling depends upon reserves provided by the parent sporophyte. In angiosperms these reserves are found in the endosperm, residual tissues of the ovule, or in the body of the embryo, usually in the cotyledons. In gymnosperms, food materials are contained mainly in the female gametophyte. Since reserve materials are partly in insoluble form—as starch grains, protein granules, lipid droplets, and the like—much of the early metabolism of the seedling is concerned with mobilizing these materials and delivering, or translocating, the products to active areas. Reserves outside the embryo are digested by enzymes secreted by the embryo and, in some instances, also by special cells of the endosperm.

In some seeds (*e.g.*, castor beans) absorption of nutrients from reserves is through the cotyledons, which later expand in the light to become the first organs active in photosynthesis. When the reserves are stored in the cotyledons themselves, these organs may shrink after germination and die or develop chlorophyll and become photosynthetic.

Orientation of the seedling

Environmental factors play an important part not only in determining the orientation of the seedling during its establishment as a rooted plant but also in controlling some aspects of its development. The response of the seedling to gravity is important. The radicle, which normally grows downward into the soil, is said to be positively geotropic. The young shoot, or plumule, is said to be negatively geotropic, because it moves away from the soil; it rises by the extension of either the hypocotyl, the region between the radicle and the cotyledons, or the epicotyl, the segment above the level of the cotyledons. If the hypocotyl is extended, the cotyledons are carried out of the soil, but, if the epicotyl elongates, the cotyledons remain in the soil.

Light affects both the orientation of the seedling and its form. When a seed germinates below the soil surface, the plumule may emerge bent over, thus protecting its delicate tip, only to straighten out when exposed to light (the curvature is retained if the shoot emerges into darkness). Correspondingly, the young leaves of the plumule in such plants as the bean do not expand and become green except after exposure to light. These adaptative responses are known to be governed by reactions in which the light-sensitive pigment phytochrome plays a part. In most seedlings, the shoot shows a strong attraction to light, or a positive phototropism, which is most evident when the source of light is from one direction. Combined with the response to gravity, this positive phototropism maximizes the likelihood that the aerial parts of the plant will reach the environment most favourable for photosynthesis.

Later development: the sporophyte plant body

CONTINUATION OF ORGAN FORMATION

Although it is convenient to refer to the early development of the plant sporophyte from the fertilized egg as embryogenesis, the process is never actually concluded as it is in the higher animals. In vascular plants, organ formation (organogenesis) is not confined to early life, and the processes of shoot, root, and leaf formation that occur first in the embryo are repeated, albeit in modified form, throughout the life of the plant. The life-span may be short and determinate, as in annual plants such as the cereals, or long, lasting for many years—indeed potentially indefinitely, except for limitations imposed by the environment and accidents—as in trees. The protracted growth of perennials, or plants that resume growth each growing season, tends to lead to increase in size, but bulk is not necessarily directly correlated with age, because individual leaves, flowers, and even whole limbs continuously die and are shed. Some long-lived plants, however, do reach a point at which losses of body mass balance the increase resulting from continued growth and organ formation.

The activity of meristems. Characteristically, vascular plants grow and develop through the activity of organ-forming regions, the growing points. The mechanical support and additional conductive pathways needed by increased bulk are provided by the enlargement of the older parts of the shoot and root axes. New cells are added through the activity of special tissues called meristems, the cells of which are small, intensely active metabolically, and densely packed with organelles and membranes, but usually lacking the fluid-filled sacs called vacuoles. Meristems may be classified according to their location in the plant and their special functions. One important distinction is between persistent meristems, typified by those of the growing points, and meristems with a limited life, those associated with organs, such as the leaf, of determinate growth. The regions of rapid cell division at the tips (apices) of the stem and the root are terminal meristems. In the stem apex, the uppermost part is the promeristem, below which is a zone of transversely oriented early cell walls, the file, or rib, meristem. The procambium is a meristematic tissue concerned with providing the primary tissues of the vascular system; the cambium proper is the continuous cylinder of meristematic cells responsible for producing the new vascular tissues in mature stems and roots. The cork cambium, or phellogen, produces the protective outer layers of the bark.

Classification of meristems

Among meristems of limited existence is the marginal, or plate, meristem responsible for the increase in surface area of a leaf; it contributes new cells mainly in one plane. Another type of meristem of limited life is called intercalary; it is responsible for the extension of some stems (as in the grasses) by the addition of new tissues remote from the growing points.

The number of dividing cells in persistent meristems remains roughly constant, with one of the daughter cells of each division remaining meristematic and the other differentiating as a component of a developing organ. The geometrical arrangements in the particular organ determine the way in which this occurs, but in general the consequence is that the meristem is continuously moving away from the maturing tissue as growth continues. It remains, therefore, a localized zone of specialized tissue, never becoming diluted by the interposition of expanding or differentiating cells. In organs such as leaves, flowers, and fruits, in which the growth is determinate, the divisions of meristematic cells become more widely scattered, and the frequency progressively falls as the proportion of the daughter cells that differentiate increases. Ultimately, at maturity, no localized meristem remains.

The contribution of cells and tissues. The two major factors determining the forms of plant tissues and organs are the orientation of the planes of cell division and the shapes assumed by the cells as they enlarge. Clearly, if the division planes in a cell mass are randomly oriented and individual cells expand uniformly, the tissue will enlarge as a sphere. On the other hand, if cell division planes are oriented regularly or the expansion of individual cells is directional, the tissue can assume any of a number of shapes. In a stem, for example, the cell division planes of the promeristem are oriented at various angles to the stem axis, so that new cells produced contribute to both width and length. Below this region, in the rib meristem, the proportion of divisions with the cell plate at right angles to

the axis increases, so that the cells tend to be oriented in files. The cells in these files expand vertically more than they do horizontally, and, accordingly, the stem develops as a cylinder.

Influence of the mitotic spindle

The factors that control the orientation of cell division planes in meristems are largely unknown. Cell interactions, however, are presumed to coordinate the distribution and orientation of the divisions. In each cell microtubules in the cytoplasm help to orient the nucleus before it divides. Then, at the time of the division, other microtubules arranged in a spindle-shaped figure (the mitotic spindle) are involved in separating the daughter chromosomes and moving them to opposite ends of the parent cell. Thereafter, the residual part of the spindle helps to locate the plate that separates the two daughter cells. Microtubules are also concerned in determining the direction of growth in expanding cells, since they appear to influence the construction of the cell wall by controlling the way cellulose is laid down in it.

Although change in shape is a form of cell differentiation, the term in the more general sense refers to a change in function, usually accompanied by specialization and the loss of the capacity for further division. Biochemical differentiation often involves a change in the character of the cell organelles—as when a generalized potential pigment body (proplastid) matures as a chloroplast, a chlorophyll-containing plastid. But it may also involve structural changes at a subcellular level, as when organelles change their character in cells engaged in intense metabolic activity.

Plant cell walls

The differentiation of plant cells for the movement of materials and the provision of mechanical support or protection invariably depends upon modification of the walls. This usually entails the accretion of new kinds of wall materials, such as lignin in woody tissue and cutin and suberin in epidermal tissues and cork. The accompanying structural changes must be controlled, for the wall materials are not applied at random but according to a pattern appropriate to the particular cell or tissue. The development of patterns during cell-wall growth depends not only on the cytoplasmic microtubules, as in the construction of the cells that will give rise to the water-conducting vessels (xylem elements), but also on cytoplasmic membranes, as in the formation of sievelike end walls (sieve plates) in the cells that will give rise to food-conducting vessels (phloem elements).

The differentiation of xylem culminates in the death of the participating cells, and the vessels are formed of chains of empty walls. This is an example of "programmed death," not an uncommon phenomenon in plant and animal development.

THE SHOOT SYSTEM AND ITS DERIVATIVES

The shoot tip. The gametophytes of mosses and liverworts and the sporophytes of many higher plants have a shoot, or early stem, with a single cell at its tip, or apex, from which all the tissues of the stem arise. This apical cell is usually four-sided (tetrahedral), with three faces directed downward, and the fourth capping the apex. Daughter cells are continually cut off sequentially from the three inner faces, the apical cell preserving its tetrahedral shape. In cell lineages derived from the daughter cells, the division planes may remain oriented in a more or less regular manner, so that, for some distance below the apex, the three sectors can be recognized in the stem. This basic pattern occurs in the arrangement of the "leaves" of some mosses, which lie in three ranks. In many plants, however, division planes in the lower part of the apex show no particular correlation with the planes of cleavage of the apical cell, and the lateral appendages do not reflect any three-part arrangement.

The generative role of promeristem tissue

Gymnosperm and angiosperm apices do not possess apical cells. The generative role is discharged by an ill-defined zone of tissue called the promeristem. Regularities may appear in the distribution of division planes only in the extreme tip region. Over the outer part of the apex, the cells often appear to lie in one to three layers, which constitute the tunica (Figure 9). Enclosed by the tunica lies a core of cells that exhibits no distinct layering; this zone is the corpus. The layers of the tunica normally contribute to the surface layers of the plant, and the corpus provides the deeper lying tissues.

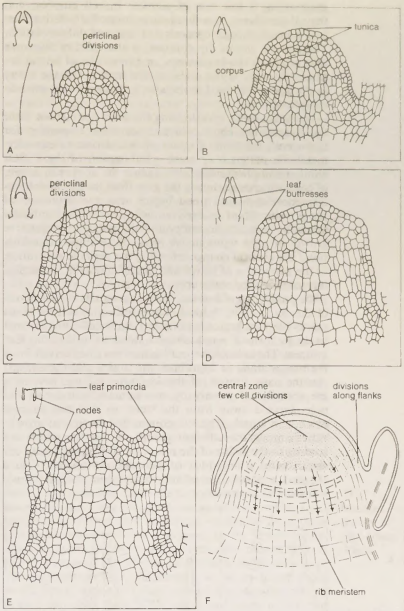


Figure 9: Leaf initiation in shoot tip of *Hypericum uralum*.
From (A–E) Zimmerman in K. Esau, *Plant Anatomy* (1965), John Wiley & Sons, Inc.; (F) F.C. Steward and M. Ram, *Advances in Morphogenesis*, 1:245, Academic Press, Inc.

The tunica–corpus analysis emphasizes the orientation of division planes, but apices can be examined from other points of view—the sizes of cells, the degree of vacuolation, and the concentration of various cell constituents, especially ribonucleic acid (RNA), vary through the apex and this sometimes results in more or less distinctive zones. Both gymnosperm and angiosperm apices have been classified on the bases of such zonal patterns, but the validity of this approach, as well as its usefulness for understanding the function of the apex as a morphogenetic centre, has been questioned.

Since 1950, a theory of angiosperm apical zonation developed by French and Belgian botanists has been gaining support. This theory proposes that the central region of the apical dome constitutes a mass of cells with relatively low division rates, the *méristème d'attente*, or "waiting meristem." Surrounding this region is an annular zone of cells with higher division rates, the *anneau initial*, or "initiating ring." Features other than division rates characterize these zones: RNA and protein content are lower in the *méristème d'attente* than in the *anneau initial*, and the nucleoli are smaller. In longitudinal section, the differences contribute to the patterns distinguishable in apices, some of which have been used as bases for structural classification. The main contention of the Franco-Belgian school, however, is that the zonation represents a functional difference. The *méristème d'attente* is regarded as a region mainly concerned with controlling the geometry of the apex. The cells have a restricted metabolism concerned primarily in

maintaining a low rate of increase in cell number, and they themselves, as well as their immediate derivatives, take no part in organogenesis or associated differentiation. The *anneau initial*, by contrast, is that part of the apex that produces the beginnings, or primordia, of lateral organs. Not only is the division rate higher, but the tissue as a whole is involved in metabolic syntheses that precede morphogenesis.

One difficulty in investigating the stem apex arises from the uncertainty about which aspects are important for the overall function: division planes, division frequency, metabolic patterns, or some combination of these. Still another complication results because the apex is in a state of constant change during the growth of the plant. A long-term developmental trend begins after the definition of the growing point in early embryogenesis and continues thereafter through juvenility and the period of vegetative growth into the reproductive phase. Superimposed on this trend is a cyclical change reflecting the periodic generation of the primordia of leaves and lateral shoots in the region immediately under the apex.

The production of leaves. Leaves originate on the flanks of the shoot apex. A local concentration of cell divisions marks the very beginning of a leaf; these cells then enlarge so as to form a nipple-shaped structure called the leaf buttress. The cells of the leaf buttress may be derived from the tunica alone or from both the tunica and the corpus.

In the early growth of the leaf primordium, new cells are contributed mainly by meristematic activity at the pole directed away from the stem, so that the buttress extends in length. The subsequent distribution of growth varies among the different groups of vascular plants according to the shape of the mature leaf. In considering the angiosperms, a broad-leaved dicotyledon (tobacco; Figure 10) and a narrow-leaved monocotyledon (maize [corn]) will serve as examples.

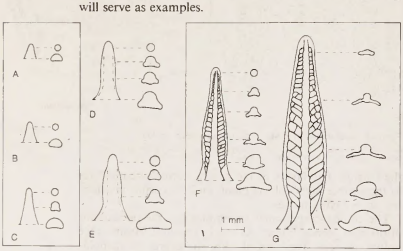


Figure 10: *Longitudinal and transverse sections of tobacco foliage leaf primordia in successive stages of development.* (A,B) The primordium in early development. (C) The small lateral ridges appearing about midway on the primordium are forerunners of thelamina. (D) Development of the lamina started. (E) Further expansion of the lamina and the beginning of development of the main lateral veins. (F) Primordium five millimetres in length showing development toward the base, of the network of provascular strands. (G) Primordium with the development of the lamina and system of venation well started. Dotted lines in C–E represent external boundaries of midrib and lateral veins and not the provascular portion.

The tobacco leaf as an example

Apical growth dominates in the tobacco-leaf primordium until a height of about 0.5 millimetre (0.02 inch) is reached. Thereafter, the buttress becomes more and more flattened in the transverse plane by laterally oriented cell divisions and further expansion growth on either side. The dividing zones are the marginal meristems, through the activity of which the leaf gains its laminate form. In each meristem the outer file of cells, or marginal initials, contributes the epidermal layers by continued division. The cells below, the submarginal initials, provide the tissue of the inner part of the leaf. Usually a certain number of cell layers is defined in the mesophyll (the parenchyma between the epidermal layers of a foliage leaf). Cell division is not limited to the region of the marginal meristems but continues throughout the leaf in each of the layers, always in the same plane, until the final cell number is approached. The rate then declines, ceasing in the different layers at different times. Divisions usually end first in the epidermis, then in the lower mesophyll layers of a leaf such as that of tobacco, and last in the main photosynthetic tissue, the palisade layer, just beneath the upper epidermis.

The vascular pattern in a tobacco leaf is determined early in the development of the vessel primordium. A procambial strand is formed by the elongation of narrow axial cells, and this extends both toward the base and toward the apex, eventually linking with the procambium of the stem. When the marginal meristems become active, the lateral veins of the leaf are initiated first, followed by the third and later order branchings that give the characteristic network of veins in the mature leaf.

Although the differentiation of the cells of the vascular system begins at the base, the epidermal and mesophyll cells mature from the tip inward toward the stem. The palisade cells elongate in a plane at right angles to the epidermis; those of the lower mesophyll expand irregularly to give lobed forms. The cells of the epidermis, shaped like irregular paving stones, continue to expand in the plane of the leaf after growth ceases in the mesophyll, so that the cells of the internal tissues are pulled apart to form the system of air spaces found in the mature leaf.

Dicotyledonous leaves are folded in various ways in the bud, the patterns being determined by differential growth in the tissues of the upper and lower surfaces (laminae) of the young leaves. Differential growth may cause the lamina either to roll or fold toward the leaf midrib or to fold near lateral veins, thus pleating the lamina. The folds in the bud are, of course, eliminated during the final phase of leaf expansion.

In the development of the maize leaf, the primordium arises first as a prominence some distance below the apical dome. The zone of division and growth extends laterally around the apex so that a complete collar forms; then the margins overlap. Meanwhile the original tip zone continues to elongate, eventually surpassing the stem apex. Tip growth declines thereafter, and further increase in cell number results from meristematic activity at the base. The early development of the vascular system is unlike that in dicotyledons, for several parallel procambial strands, rather than a single midrib, are initiated. The first of these grow toward the apex, but, as tip growth ceases, procambial strands form above and extend toward the base, passing through the node, or point of insertion of the leaf primordium, and into the stem below. As the leaf extends in length, the tissues begin to mature first at the tip, and a wave of differentiation passes down toward the base, where cell division and extension growth may continue long after the tip of the leaf is mature. Protection for this immature and succulent tissue of the leaf base is afforded by the sheaths of older leaves surrounding it.

Variations in leaves

These examples illustrate the principles involved in leaf development, but there are many deviations associated with variations in leaf form. Lobing and toothing result from the persistence of cell division and growth in particular stretches of the margin after growth ceases in between. Carried to the extreme, this localized growth gives the feather-like pinnate leaf. Many monocotyledons form cylindrical leaves as a result of a fusion of the margins of the primordium after it has encircled the stem.

Branching of the shoot. The shoots of most vascular plants branch according to a consistent plan, with each new axis arising in the angle between a leaf and a stem—that is, in a leaf axil. In some plants, buds may also form from the older parts of shoot or root remote from the main apices; these buds, termed adventitious, do not conform to the general plan.

A lateral shoot apex is initiated on the flanks of the main apex but at some distance below the point of emergence of the youngest leaf primordium. As in the origin of a leaf, generally the outer cell layers contribute to the surface tissues of the new apex by maintaining a consistent pattern of divisions. In some species a tunica of more than one cell layer quickly forms, so that the new apex appears as a miniature version of the main one; alternatively, the differentiation may not become apparent until the new

primordium has attained considerable bulk. In all cases, the new apex must reach a minimal volume before it in turn can begin to form its own lateral primordia and to organize true axillary buds. As this volume is attained, *méristème d'attente–anneau initial* zonation appears. As in the main apex, the formation of new primordia is associated with the annular zone.

From this point on, the development of the lateral shoot is the same as that of the main shoot, except that growth may not be as rapid because the main apex, or leading bud, dominates and absorbs much of the available nutrient. The early growth of the axillary bud proceeds quite vigorously until a certain number of leaf primordia has been formed; then apical activity slows. Cell division gradually stops, and with it the associated syntheses; thus there is no increase in the DNA of the nuclei of the meristem after the last division. The bud, in effect, passes into a state of dormancy, even though the external conditions for growth are propitious. This phenomenon is known as correlative bud inhibition, since it is determined by the activity of the leading bud of the shoot. If the leading bud is removed, the inhibited lateral buds resume growth, and with it the associated syntheses.

Vascular development. Cell division planes in the zone just below the apex of the shoot tend to be oriented so that vertical files of cells are formed. This is more evident in the central core than in the surrounding cortical region, for the pattern is not disturbed by the insertion of lateral members. The first signs of the differentiation of the vascular system appear some distance below the apex, in a zone of tissue distinguishable by the smaller cross-sectional area of individual cells. These cells, forming the procambium zone, arise by divisions oriented at right angles to the axis and may form a complete cylinder; generally, however, interruptions occur, the segments being related to the uppermost leaf primordia. In a dicotyledon such as tobacco, the cylinder at its highest level consists of strands running upward toward the points of insertion of the primordia. Thus, as the site of each primordium is determined, a strand forming in the adjacent region of the stem will contribute to the cylinder, but at a higher level than the preceding strand. The link with the earlier formed procambium is not simple, however. The strand passing upward toward a leaf primordium usually is composed of branches arising from strands that enter the two nearest older leaves below it. Because it in turn will contribute a branch for the next leaf, the cylinder is really a hollow network, the "gaps," or leaf traces, marking the points of departure of the leaf veins.

The "cylinder" of vessels

During subsequent development, the strands, or vascular bundles, increase in thickness by further cell divisions, and connections form with the vascular systems of axillary buds. The cells differentiate to give the characteristic tissues of the vascular system: phloem vessels (conducting tissue), phloem parenchyma (packing tissue), and phloem fibres (supporting tissue) toward the outside and xylem vessels (woody conducting tissue) toward the inside. The differentiation occurs in an upward direction, so that the maturation of the vascular tissues follows at a more or less constant distance behind the apex.

Although details differ, the above account of the origin of the primary vascular system is broadly applicable to gymnosperms and many ferns. Vascular development differs somewhat in certain flowering plants. In many monocotyledons, such as maize, the several vascular strands that pass down from each leaf primordium into the stem do not contribute to a single cylinder but are scattered in the ground tissue, or parenchyma, of the stem. Lateral interconnections form principally at the nodes.

Cambium activity

Increase in stem diameter is accomplished in the older stems of dicotyledons by the activity of the cambium, which produces secondary vascular tissue. This meristem, a relic of the procambium, is composed of thin-walled cells, is flattened in the radial plane, and persists between primary vascular tissue, the differentiated outer phloem, and the inner xylem. When secondary thickening begins, the parenchymatous cells between the vascular bundles also resume division, ultimately forming a cambium cylinder. The cells of the cambium divide, producing initial phloem cells toward the outside and initial xylem cells within. Files of cells are also cut off among the initial phloem and xylem cells that remain parenchymatous and are called phloem and xylem rays.

As in the apical meristem, the number of dividing cells in the cambium remains constant, except that more cells occasionally are added by divisions in the radial plane so that the girth of the cambial cylinder expands in pace with the growth of the xylem within. The addition of new phloem toward the outside compresses the primary phloem and the cortical tissues in a radial direction while stretching them tangentially. These primary tissues do not persist, however. As the girth of the stem increases, the epidermis is disrupted, and the outer layers of the cortex become meristematic, giving rise to the cork cambium, which generates cork cells on the outside. The cork layer, or bark, then takes over the protective function of the epidermis.

THE ROOT SYSTEM AND ITS DERIVATIVES

The root tip. Plants that have a single apical cell in the shoot also have a single apical cell in the root. The cell is again tetrahedral, but sometimes daughter cells are cut off from all four faces, with the face directed away from the axis producing the cells of the root cap. The cells derived from the other faces continue to divide mostly by forming transverse walls, but occasionally also in the longitudinal plane. In this way vertical columns of cells form—tending, because of their mode of origin, to be disposed in three sectors.

In the roots of gymnosperms, angiosperms, and some lower plants, there is no single apical cell. Again, as with the shoot, such root apices can be analyzed in different ways. Perhaps the most useful approach is based upon tracing the sources of the main tissues in the apical region. Such an analysis has led to the histogen theory, which proposes that the three principal tissues of the root—vascular cylinder, cortex, and epidermis—originate from three groups of initial cells, or histogens, in the apical meristem—plerome, periblem, and dermatogen respectively. A fourth histogen, the calyptrogen, produces the root cap. The histogens have been thought to lie in linear order in the apex, with the initial cells of the vascular system toward the older part of the root, and those of the cap toward the tip.

The histogen theory of root-tissue origin

The histogen theory is difficult to apply to some types of roots, and there has been uncertainty about the numbers of histogens. The discovery of the "quiescent centre" in the root apex has clarified many features, however. The quiescent centre is a group of cells, up to 1,000 in number, in the form of a hemisphere, with the flat face toward the root tip; it lies at the centre of the meristem, in much the same position, in fact, as the tetrahedral apical cell in certain lower plants (Figure 11). The cells of the quiescent centre are unusual in that their division rate is lower than that in the surrounding meristem. The cells of the centre have other distinctive features as well, notably a lower rate of protein synthesis than that of neighbouring cells.

The quiescent centre is surrounded by actively dividing cells of the promeristem that are the initial cells of the various tissues of the root. Those abutting the flat, tip-directed face contribute to the root cap; those above the quiescent centre are distributed in a cup shape. The cells in the centre of the cup produce the procambium and so, ultimately, give rise to the vascular cylinder. The annular zone of cells surrounding this central group forms the initials of the cortex; surrounding this, in turn, a ring of initial cells forms the protoderm, the layer corresponding to the epidermis at this level of the root.

The quiescent centre, a constant feature of the root tip, is apparently generally present in angiosperm and probably also in gymnosperms. The quiescent centre probably plays a role comparable with that of the apical cell in some lower plant roots, maintaining the geometry of the system. It has also been suggested that it may be concerned with the synthesis of growth hormones, although no direct evidence exists. When roots are damaged mechanically or by radiation, the cells of the centre can resume a rapid division rate, and they then participate in regeneration.

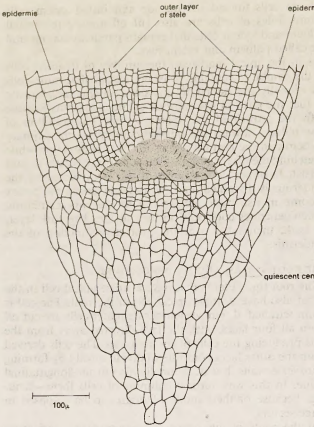


Figure 11: Median section of root apex of *Zea mays*.
Adapted from *Biological Reviews of the Cambridge Philosophical Society* (1959)

The zone of cell division extends some distance along the length of the root above the tip region. Although the girth may increase by longitudinal divisions and the widening of the daughter cells, most divisions occur in the transverse plane resulting in the formation of longitudinal files of cells.

In longitudinal section, the tissue zones become progressively better defined away from the tip. An internal protective band, the endodermis, becomes conspicuous as a single sheath of cells surrounding the procambium. The phloem procambium, recognizable by its narrow cells, begins to differentiate in the lower part of the region of elongation. The xylem also becomes distinct, the thickenings appearing first in the upper part of the extension zone. Differentiation keeps pace with the advance of the root tip as new cells are added in the promeristem. When xylem occupies the core, there is no pith as in the shoot, but the cells of the outermost layers of the vascular cylinder remain undifferentiated, forming the pericycle, a tissue important in the formation of lateral roots. Within the bounds of the pericycle, the xylem is star-shaped in section, with the first-formed xylem elements (protoxylem) occupying the ridges. The phloem lies in the intervening grooves. Outside of the endodermis, the cortical cells elongate but remain thin-walled. Above the level of the root-cap sheath, the epidermis forms the outer layer of the root, and, beyond the extension zone, its cells begin to develop root hairs.

A more complete account can be given for the mechanics of development of the root apex than for that of the stem, mainly because of its greater simplicity. An important difference lies in the absence of a mechanism for the cyclical production of lateral organs at the apex itself.

Generation of lateral roots

Branching of the root. The branching of the root takes place in the older parts and does not directly involve the apical meristem. The tissues concerned are the endodermis and the layer immediately beneath it, the pericycle. The endodermis participates in root branching in certain lower plants with apical cells. A cell of this layer enlarges and forms a tetrahedral cell, which becomes the new apical cell; by further divisions a hemispherical volume of tissue forms around it—the whole constituting a new apex.

In many other plants, including gymnosperms and angiosperms, the lateral roots develop from the pericycle. Cells in this layer enlarge and begin to divide until a dome of tissue develops. Called the incipient apex, the dome pushes out the surrounding endodermis, which may itself resume divisions, its daughter cells enlarging to create a sheath around the new root tip. During further growth, the dome assumes an organization like that of the primary root apex. At first, all cells are meristematic; then, while the primordium is still small, cells in the central zone cease DNA synthesis, and this zone becomes the new quiescent centre. Beyond it, the root cap is produced, and, at the base, initial cells begin to develop the cell files that become the vascular cylinder, cortex, and epidermis. The vascular tissues differentiate from the base outward, and link eventually with xylem and phloem of the parent root. All this development occurs before the tip of the new root emerges from the tissues of the parent root. The growth of the new tip into the cortex first pushes out the endodermal sheath, if one is present, and then bursts it. The cortical cells are themselves crushed and probably resorbed as the root grows on, until finally the tip breaks through the epidermis.

In most roots, new laterals are initiated in the pericycle opposite to the protoxylem ridges. They tend accordingly to form vertical ranks along the length of the root, reflecting the number of bands of protoxylem. Although lateral roots arise in quite a different way from leaves and axillary shoots at the stem apex, there are certain common features. Pericyclic cells about to produce a root primordium synthesize ribonucleic acid, in anticipation of the period of growth and morphogenesis that will result in a new apex. The same behaviour is seen in the cells of the annular zone, from which leaf primordia arise at the stem apex, and also in the axillary zones at a slightly lower level, from which new stem apices develop.

Later growth. In the secondary growth of the root, cell division in the primary xylem produces a cambium, which abuts the pericycle over the protoxylem ridges and passes between the phloem strands and the xylem in the grooves. Activity of the root cambium is comparable with that of the stem cambium; phloem elements are cut off outward, and xylem elements are cut off within. With continued growth in thickness, the star-shaped figure of the primary xylem is lost, and the cambium eventually forms a cylindrical sheath. Again, as in the stem, the protective function of the epidermis is ultimately taken over by cork layers produced by a cork cambium in the outer cortex.

Correlations in plant development

COORDINATION OF SHOOT AND ROOT DEVELOPMENT

Although the structural organization of the vascular plant is comparatively loose, development of the various parts is well coordinated. Control is dependent upon the movement of chemical substances, including both nutrients and hormones.

Meeting demands for food and support

An example of correlation is the growth of shoot and root. The enlargement of aerial parts is accompanied by increased demands for water, minerals, and mechanical support that are met by coordinated growth of the root system. Several factors apparently are concerned with control, because shoot and root affect each other reciprocally. The root depends on the shoot for organic nutrients, just as the shoot depends on the root for water and inorganic nutrients and the flow of ordinary nutrients must, therefore, play some part. More specific control, however, may be provided by the supply of nutrients required in very small amounts. The root depends on the shoot for certain vitamins, and variation in the supply, reflecting the metabolic state of the aerial parts, may also influence root growth. In addition, hormonal factors affecting cell division pass upward from the root into the stem; although the exact role of the hormones has not yet been established with certainty, they may provide one way by which the root system can influence the activity of the shoot apex.

The control of secondary thickening is another important example of growth correlation. As the size of the shoot system increases, the need for both greater mechanical support and increased transport of water, minerals, and manufactured food is met by an increase in stem girth through the activity of the vascular cambium. Generally, the cambium of trees in temperate zones is most active in

the spring, when buds open and shoots extend, creating a demand for nutrients. Cell division begins near the bud in each shoot and then spreads away from it. The terminal bud stimulates the cambium to divide rapidly through the action of two groups of plant hormones: auxins and gibberellins.

The inhibition of lateral buds, another example of correlated growth response, illustrates a reaction opposite to that occurring in the control of cambial activity. Lateral buds are inhibited in general because axillary shoots grow more slowly or not at all, while the terminal bud is active. This so-called apical dominance is responsible for the characteristic single trunk growth seen in many conifers and in herbaceous plants such as the hollyhock. Weaker dominance results in a bushy growth form with repeated branching. The fact that lateral, or axillary, buds become more active when the terminal bud is removed suggests that hormonal control is involved.

The flow of auxin from the shoot tip is, in part, responsible for inhibiting axillary buds. The nutritional status of the plant also plays a role, apical dominance being strongest when mineral supply and light are inadequate. Because axillary buds are released from inhibition when treated with cell-division promoting substances (cytokinins), it has been suggested that these substances are also concerned in regulating axillary-bud activity.

DETERMINATION OF MATURE FORM

The duration of a juvenile phase

After its establishment as an independent plant, the sporophyte passes through a juvenile period before reaching maturity and becoming reproductive. Juvenility may be brief or, as in the case of trees, may extend over several years. The duration is determined partly by internal factors and partly by environmental controls related to the seasons.

In some ways juvenility is a continuation of developmental trends initiated in the embryo. In many plants, new organs are produced sequentially through early life, each of progressively more mature form. The first leaf of the young fern sporophyte, for example, is small and relatively simple, and the vascular system consists of a few forked strands. As growth proceeds, succeeding new leaves are of increasing complexity, and the shape begins to resemble that typical of the reproductive frond; in addition, vasculation shifts to the mature pattern, often one with a network of veins. Comparable trends occur in flowering plants, in which leaves at successive levels of plant maturity often show a progressive increase in the complexity of lobing or toothing.

Some of the changes associated with the juvenile period can be attributed to the gradual enlargement of the growing point, necessarily small in the embryo; its volume increases progressively with development. This increase in cell number is usually associated with the emergence of a "mature" zonation pattern. The typical internal structure of the shoot apex does not develop until a specific number of leaves form.

Gradual structural change in the growing point, however, does not adequately account for all aspects of juvenility. Sometimes, the transition from juvenile to adult leaf form is not graded but sudden. The juvenile leaves of species of the gymnosperm *Chamaecyparis*, for example, are needlelike and spreading; the adult leaves are scalelike and lie close to the stem. Among flowering plants, various species of *Eucalyptus* have juvenile leaves that are ovate and mature leaves that are sickle-shaped.

Phase change

Such sudden transitions from juvenile to adult form, referred to as phase change, seem to depend not on slow shifts in the apex but on some determinative event or correlated group of events. The two forms are relatively stable and tend to resist change; for example, cultured tissues taken from the juvenile (ivy-leaved) parts of ivy plants maintain a higher rate of cell division, and portions, or cuttings, taken from these parts tend to form roots more readily than those from the adult (simple-leaved) parts.

The establishment of these relatively stable but not wholly irreversible states is comparable with the determination of shoot and root poles during embryogenesis and, indeed, with the alternation of generations itself. The transmission of differentiated states through cell lineages presumably reflects the action of "switching" devices controlling the expression of different parts of the genetic complement. In this sense, phase change and related phenomena do not differ essentially from those of differentiation and organogenesis in general.

The transition in plants to the reproductive state is an example of a developmental event with some of the characteristics of phase change. Among seed plants, the reproductive structures are transformed shoots—strobili (including cones) of various kinds in the gymnosperms and flowers in angiosperms.

Flower development

From a developmental point of view, the flower can be regarded as a shoot axis of determinate growth, with the lateral members occupying the sites of leaves differentiating as floral organs—sepals, petals, stamens, and pistils. In the transition to flowering, the stem apex undergoes distinctive changes, the most conspicuous of which is in the shape of the apical region, which is related to the kind of structure to be formed, whether a single flower, as in the tulip, or a cluster of flowers (an inflorescence), as in the lilac. The region of cell division extends over the entire apex, and the ribonucleic acid content of terminal cells increases. When a single flower forms, lateral primordia emerge at higher and higher levels on the flanks of the apical dome, and the entire apex is absorbed in the process, after which apical growth ceases. When an inflorescence forms, early changes are generally comparable to that for the single flower with one major difference—axillary primordia emerge that either become floral meristems or develop as secondary inflorescence branches. These primordia appear closer to the apex than do those of axillary buds on a vegetative shoot. In grasses, the activation of axillary meristems is the most notable early indication of the passage into flowering.

The rate of maturation and the timing of the transition to the reproductive phase are sometimes governed by internal controls and thus are relatively insensitive to the environment, provided conditions are generally favourable for growth. Frequently, however, the developmental rate is affected profoundly by recurring cycles in the environment, particularly those of temperature and of day length. In effect, these cycles provide a timetable for the plant, thus adjusting flowering, fruiting, and seed dispersal to the season and increasing the chances for successful propagation.

Vernalization of the seed

The control of the developmental rate by temperature is especially evident in many herbaceous plants of temperate climates. These plants, as indicated earlier, often must experience cold, either as seeds or as young plants, before they can begin flower production; otherwise they undergo an excessively long period of leafy, or vegetative, growth. After the cold experience, which can be given artificially, the plant is said to have been vernalized, or brought to the spring condition. Again the response is akin to a determination, because the condition attained is transmitted through subsequent cell divisions. Furthermore, there are indications that vernalization induces a persistent modification in the metabolism of apical cells and their derivatives. Ingenious theoretical schemes, offered to explain the apparent paradox that low temperature should actually accelerate a developmental process, are based mostly upon the proposition that a special vernalization hormone (vernalin) is involved. Although little direct evidence for the existence of vernalin exists, a class of hormones found in certain plant species, the gibberellins, does participate. The cold requirements of some species, such as the carrot, can be eliminated by the application of gibberellin, although the amounts needed are substantial.

Photoperiod as a trigger for flowering

The annual cycle of changing day length obviously provides the best of all "clocks" for the regulation of plant development. The effect of day length (or rather length of continuous darkness) on the transition to flowering is part of the general phenomenon of photoperiodism. Certain plants, called short-day plants, grow vegetatively when the nights are shorter than a critical minimum period (days long); exposure to longer nights (days short), however, accelerates development and brings on early flowering. Conversely, long-day plants develop very slowly toward flowering during daily cycles with longer than a minimum

of darkness (days short), and are accelerated by exposure to short nights (days long). Other plants either require days of intermediate length for flowering or respond to a sequence of different photoperiods.

The leaf, rather than the stem apex, is the light-receiving organ in the photoperiodic reaction, although it is at the apex that subsequent developmental changes occur. One commonly accepted view is that, as a consequence of the photoperiodic experience, a specific flower-inducing hormone (as yet not isolated but referred to as "florigen") is synthesized in the leaf and translocated to the apex. As in the case of vernalization, photoperiod undoubtedly affects the metabolism of the known plant hormones, and so influences many other developmental responses apart from flowering. The effect of the duration of illumination on the carbohydrate balance of the plant may also be important. Nutritional effects on flowering are well known in many species—certain fruit trees, for example.

Whether or not environmental factors influence the passage into a reproductive state of a plant, the transition must be viewed as part of the general development from juvenility to maturity: in this sense, flowering is not a radical alternative to vegetative growth but its culmination. Yet, entirely new organ types are produced at the flowering apex, presumably under the influence of genes inactive during vegetative growth.

SEASONAL ADAPTATIONS

Certain plants are perennial and survive from year to year by matching their growth to the progression of the seasons or by suspending growth altogether during unfavourable times, such as winter or a dry season.

Response to an approaching unfavourable season

In the temperate zone, some time before winter begins, growth ceases in the shoots of woody plants, resting buds are formed, and deciduous trees lose their leaves. The resting bud consists of a short axis, with the stem apex surrounded by modified unexpanded leaves, which protect the stem, especially from drying. The cells show marked frost resistance, similar to that of the embryo of the seed. Corresponding changes occur in herbaceous plants, in which the preparation for winter may involve the dying back of aerial parts altogether, leaving protected organs at or below the soil surface.

Growth sometimes ceases, even under favourable conditions, as a result of internal changes in the plant. This is true for some trees, which cease growth in midsummer. The passage into winter dormancy, however, is often controlled by the shortening of day length at the end of the growing season; in some plants decreasing night temperature also plays a part. Most temperate zone trees cease growth and form resting buds when the day length falls below a critical minimum.

Photoperiodic control seems to involve the formation of inhibitor compounds. In birches, for example, the leaf perceives the day length "signal" and transmits inhibitory materials to the apex, thus bringing growth to a stop and inducing the formation of a resting bud. The dormancy hormone, abscisic acid, may be concerned in this response and also in leaf abscission.

Budbreak

Budbreak in certain trees is controlled by photoperiod, growth resuming in the lengthening days of spring; light-perceptive organs are probably the young leaves inside the bud scales. Sometimes budbreak depends only on temperature increases that occur in spring, as in certain plants of Mediterranean climates.

The resumption of development in buds may result from a change in the balance of growth-inhibiting substances, such as abscisic acid, and growth promoters, notably the gibberellins. Buds can be caused to open prematurely by gibberellin treatment, which, as in the case of vernalization, can sometimes replace a cold experience; moreover, the gibberellin content in the buds of certain woody plants increases during chilling. Other hormones are probably also involved, however, for budbreak in plants such as the grapevine can be promoted by cytokinins, the plant cell-division factors.

An important general feature of adaptive periodicities is that the developmental changes anticipate the conditions for which they will ultimately provide the appropriate physiological or morphological adjustment. The ability of plants to utilize environmental indicators such as temperature and day-length changes is vital for the survival of plants. The production of such adaptive devices is made possible by the state of continuous embryogeny, already stressed as one of the most important characteristics of plant growth. (J.H.-H.)

ANIMAL DEVELOPMENT

General features

Development, in the context of this section, includes the processes that lead eventually to the formation of a new animal starting from cells derived from one or more parent individuals. Development thus occurs following the process by which a new generation of organisms is produced by the parent generation.

REPRODUCTION AND DEVELOPMENT

Metazoan reproduction

In multicellular animals (Metazoa), reproduction takes one of two essentially different forms: sexual and asexual. In asexual reproduction the new individual is derived from a blastema, a group of cells from the parent body, sometimes, as in *Hydra* and other coelenterates, in the form of a "bud" on the body surface. In sponges and bryozoans, the cell groups from which new individuals develop are formed internally and may be surrounded by protective shells; these bodies, which may serve as resistant forms capable of withstanding unfavourable environmental conditions, are released after the death of the parent. In certain animals the parent may split in half, as in some worms, in which an individual worm breaks into two fairly equal parts (except that the anterior half receives the mouth, "brain," and sense organs if they are present).

Obviously, in such a case it is impossible to say which of the two resulting individuals is the parent and which the offspring. Some brittle stars (starfish relatives) may reproduce by breaking across the middle of the body disk, with each of the halves subsequently growing its missing half and the corresponding arms.

A common feature of all forms of asexual reproduction is that the cells—always a substantial number of cells, never only one cell—taking part in the formation of the new individual are not essentially different from other body, or somatic, cells. The number of chromosomes (bodies carrying the hereditary material) in the cells participating in the formation of a blastema is the same as in the other somatic cells of the parent, constituting a normal, double, or diploid ($2n$), set.

In sexual reproduction, a new individual is produced not by somatic cells of the parent but by sex cells, or gametes, which differ essentially from somatic cells in having undergone meiosis, a process in which the number of chromosomes is reduced to one-half of the diploid ($2n$) number found in somatic cells; cells containing one set of chromosomes are said to be haploid (n). The resulting sex cells thus receive only half the number of chromosomes present in the somatic cell. Furthermore, the sex cells are generally capable of developing into a new individual only after two have united in a process called fertilization (see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *The process of fertilization*).

Advantages of sexual and asexual reproduction

Each type of reproduction—asexual and sexual—has advantages for the species. Asexual reproduction is, at least in some cases, the faster process, leading most rapidly to the development of large numbers of individuals. Males and females are independently capable of producing offspring. The large size of the original mass of living matter and its high degree of organization—the new individual inherits parts of the body of the parent: a part of the alimentary canal, for instance—make subsequent development more

simple, and the attainment of a stage capable of self-support easier. New individuals produced by asexual reproduction have the same genetic constitution (genotype) as their parent and constitute what is called a clone. Though asexual reproduction is advantageous in that, if the parent animal is well adapted to its environment and the latter is stable, then all offspring will benefit, it is disadvantageous in that the fixed genotype not only makes any change in offspring impossible, should the environment change, but also prevents the acquisition of new characteristics, as part of an evolutionary process. Sexual reproduction, on the other hand, provides possibilities for variation among offspring and thus assists evolution by allowing new pairs of genes to combine in offspring. Since all body cells are derived from the fertilized egg cell, a mutation, or change, occurring in the sex cells of the parents immediately provides a new genotype in each cell of the offspring. In the course of evolution, sexual reproduction has been selected for, and established in, all main lines of organisms; asexual reproduction is found only in special cases and restricted groups of organisms.

PREPARATORY EVENTS

In the case of multicellular animals we find there are two kinds of sex cells: the female sex cell (ovum, or egg), derived from an oocyte (immature egg), and the male sex cell (spermatozoon or sperm), derived from a spermatocyte. Eggs are produced in ovaries; sperm, in testes. Both the egg and the sperm contribute to the development of the new individual; each providing one set of genes, thereby restoring the diploid number of chromosomes in the fertilized egg. The sperm possesses a whiplike tail (flagellum) that enables it to swim to the egg to fertilize it. In most cases the egg, a stationary, spherical cell, provides the potential offspring with a store of food materials, or yolk, for its early development (Figure 12). The term yolk does not

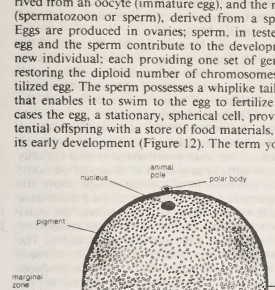


Figure 12: Amphibian egg, vertical section, showing distribution of yolk and pigment in interior.

refer to any particular substance but in fact includes proteins, phosphoproteins, lipids, cholesterol, and fats, all of which substances occur in various proportions in the eggs of different animals. In addition to yolk, eggs accumulate other components and acquire the structure necessary for the development of the new individual. In particular the egg acquires polarity—that is, the two ends, or poles, of the egg become distinctive from each other. At one pole, known as the animal pole, the cytoplasm appears to be more active and contains the nucleus (meiotic divisions occur in this region); at the other, called the vegetal pole, the cytoplasm is less active and contains most of the yolk. The general organization of the future animal is closely related to the polarity of the egg.

When the amount of food reserve is comparatively small, as it is in many marine invertebrates and mammals (in the latter the embryo is nourished by materials in the mother's blood), the egg may be barely visible to the unaided eye. The egg of the sea urchin is about 75 microns (0.003 inch) in diameter; that of a human being is slightly more than 0.1 millimetre. Eggs are classified according to the amount of yolk present. An egg with a small quantity of evenly distributed yolk is called an oligolecithal egg. One with more yolk that is unevenly distributed (*i.e.*, concentrated towards the vegetal pole) is telolecithal; and one with still greater amounts of yolk in granules or in a compact mass is megalecithal.

Protective membranes

The egg is surrounded by protective membranes, which may be soft and jellylike or hard and calcified, like shells. Egg membranes are produced while the egg is either in the ovary or being carried away from the ovary in a tube called an oviduct. The eggs of many animals have both kinds of membranes. In insects, a hard shell (chorion) forms around the eggs in the ovaries. In frogs, a very thin vitelline membrane forms around the eggs in the ovary; subsequently a layer of jelly is deposited around the eggs while they pass through the oviducts. In birds, a very thin vitelline membrane is produced around the egg in the ovary; then several layers of secondary membranes are formed in the oviduct before the egg is laid. The outermost of these secondary membranes is the calcareous shell. In mammals the egg is surrounded by the so-called pellucid zone, which is equivalent to the vitelline membrane of other animals; follicle cells form an area called the corona radiata around this zone.

After fertilization the egg, now called a zygote, is endowed with genes from two parents and has begun actual development. (Activation of the egg may be brought about by an agent other than sperm in certain animals, but such cases of parthenogenesis are exceptional. (See REPRODUCTION AND REPRODUCTIVE SYSTEMS.)

After fertilization, the zygote undergoes a series of transformations that bring it closer to the essential organization of the parents. These transformations, initiated at a physiological, perhaps even at a molecular, level, eventually result in the appearance of certain structures. The whole process is called morphogenesis (Greek *morphē*, "shape" or "form"; *genesis*, "origin" or "production"). The process of development is more easily understood if, at every step, the changes necessary to bring the system nearer the goal are considered. Depending on the achievements necessary at any step, development can be subdivided into a number of discrete phases, the first of which, cleavage, immediately follows fertilization.

Early development

EMBRYO FORMATION

Cleavage. Since the goal of development is the production of a multicellular organism, many cells must be produced from the single-celled zygote. This task is accomplished by cleavage, a series of consecutive cell divisions. Cells produced during cleavage are called blastomeres. The divisions are mitotic—*i.e.*, each chromosome in the nucleus splits into two daughter chromosomes, so that the two daughter blastomeres retain the diploid number of chromosomes. During cleavage, almost no growth occurs between consecutive divisions, and the total volume of living matter does not change substantially; as a consequence, the size of the cells is reduced by almost half at each division. At the beginning of cleavage, cell divisions tend to occur at the same time in all blastomeres, and the number of cells is doubled at each division. As cleavage progresses, the cells no longer divide at the same time.

Cleavage in most animals follows an orderly pattern, with the first division being in the plane of the main axis of the egg (Figure 13). This cleavage plane is arbitrarily called vertical, on the assumption that the main axis of the egg is vertical. The second cleavage plane is again vertical but at right angles to the first, giving rise to four equal cells arranged around the main axis of the egg (Figure 13). The third cleavage plane is at right angles to both the first and second cleavage planes and is horizontal, or equatorial. Subsequent divisions may alternate between vertical and horizontal cleavage planes, but later cleavage divisions become randomly oriented. This pattern is typical of many animal groups; however, more complicated patterns of cleavage are found in such animals as annelids, mollusks, and nematodes.

As the amount of yolk in the egg increases, it influences cleavage by hindering the cytoplasmic movements

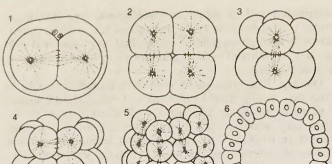


Figure 13: Cleavage of an oligolecithal egg (sea cucumber) in progressive stages 1 through 6 (see text).

Influence of yolk on cleavage

involved in mitosis. If there is only little yolk (oligolecithal eggs), the yolk granules follow the movements of the cytoplasm and are distributed in the resulting blastomeres. But if the amount of yolk is larger (megalecithal eggs), cleavages occur nearer the animal pole, where there is less yolk; as a result, the blastomeres nearer the animal pole are smaller than those nearer the vegetal pole. The presence of yolk masses may retard the onset of cleavage in a part of the egg or even suppress it altogether; in this case cleavage is partial, or meroblastic. Only a part of the egg material then is subdivided into cells, the rest remaining as a mass that serves as nourishment for the developing embryo.

Cleavage is complete, or holoblastic, in many invertebrates including coelenterates, annelids, echinoderms, tunicates, and cephalochordates. The blastomeres may be either about equal or only slightly different in size. Cleavage in amphibians is holoblastic, but the size of the blastomeres is very uneven. Blastomeres are smallest at the animal pole and largest (and yolky) at the vegetal pole. Somewhat similar conditions prevail in many mollusks. In most fishes, birds, reptiles, and egg-laying mammals (monotremes), cleavage is discoidal—*i.e.*, restricted to a disk of cytoplasm at the animal pole of the egg, most of the yolky egg material remaining uncleaved (Figure 14). Cleavage in insects and many other arthropods is

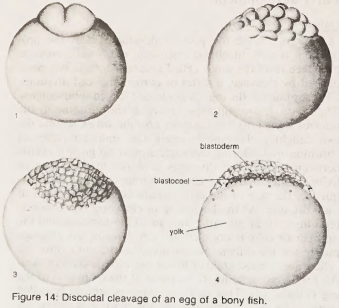


Figure 14: Discoidal cleavage of an egg of a bony fish.

superficial—*i.e.*, the entire surface layer of egg cytoplasm subdivides into cells, and the egg contains a central mass of uncleaved yolk. The conditions of cleavage in placental mammals, including man, are peculiar.

During cleavage, development involves only an increase of cell numbers; the shape of the embryo does not change, and chemical transformations within the embryo are restricted to those necessary for cell division. Chemical and structural transformations are concerned with accumulating chromosomal material in the nuclei of the blastomeres.

Chromosome doubling

Before each division the chromosomes carrying the genes double in number; this means that the chromosomal material, deoxyribonucleic acid (DNA), has to be synthesized. This synthesis proceeds possibly at the expense of cytoplasmic ribonucleic acid (RNA) but certainly also from simpler organic compounds. A certain amount of protein synthesis is also necessary for cleavage to proceed: if developing eggs are treated with puromycin, a substance which is known to suppress protein synthesis, cleavage stops immediately. The proteins concerned have not yet been identified. No proteins are synthesized, however, that would foreshadow the future differentiation of parts of the embryo. It is believed that the genes in the chromosomes remain largely inactive during cleavage. The rhythm (speed) of cleavage is wholly dependent on the cytoplasm of the egg.

Although the shape and volume of the embryo do not change during cleavage, one important change in gross organization does take place. As the blastomeres are produced, they move outward, leaving a centrally located fluid-filled cavity. In cases of holoblastic cleavage, the blastomeres become arranged in a layer from one to several cells thick surrounding the cavity. The embryo at this stage may be likened to a hollow ball and is known as a blastula (Figure 15:1). The outer layer of cells is called the blastoderm, and the fluid-filled cavity the blastocoel. In discoidal cleavage the cells, which do not surround the whole embryo, lie only on the animal pole; nevertheless, a blastocoel may be formed by a crevice appearing between the blastomeres and the mass of yolk (Figure 14). The blastomeres then may be arranged as a saucer-shaped blastodisk covering the blastocoel.

The formation of the blastula signifies the end of the period of cleavage. The next stage of development is concerned not with an increase in cell number, though cell divisions continue at a slower pace, but with rearrangement of the available cell masses to conform with the gross features of the future animal.

Gastrulation. The embryo in the blastula stage must go through profound transformations before it can approach adult organization. An adult multicellular animal typically possesses a concentric arrangement of tissues of the body; this feature is common to all animal groups above the level of the sponges. Adult tissues are derived from three embryonic cell layers called germinal layers: the outer layer is the ectoderm, the middle layer is the mesoderm, and the innermost layer is the endoderm (entoderm). The ectoderm gives rise to the skin covering, to the nervous system, and to the sense organs. The mesoderm produces the muscles, excretory organs, circulatory organs, sex organs (gonads), and internal skeleton. The endoderm lines the alimentary canal and gives rise to the organs associated with digestion and, in chordates, with breathing.

Germinal layers

The blastula, which consists of only one cell layer, undergoes a dramatic reshuffling of blastomeres preparatory to the development of the various organ systems of the animal's body. This is achieved by the process of gastrulation, which is essentially a shifting or moving of the cell material of the embryo so that the three germinal layers are aligned in their correct positions.

The rearrangement of the blastula to form the germinal layers is seen clearly in certain marine animals with oligolecithal eggs. The hollow blastula consists of a simple epithelial layer (the blastoderm), the transformation of which can be likened to the pushing in of one side of a rubber ball (Figure 15). As a result of such inpushing (or invagination), the spherical embryo is converted into a double-walled cup, the opening of which represents the position of the former vegetal pole. The involuted part of the blastoderm, lining the inside of the double-walled cup, gives rise to the endoderm and mesoderm, and the blastomeres remaining on the exterior become the ectoderm. As a consequence of the infolding at the vegetal pole, the blastocoel is reduced or obliterated, and a new cavity is created, the primitive gut cavity, or archenteron, which eventually gives rise to the hollow core (lumen) of the alimentary canal. At this stage the embryo has a primitive gut with an opening to the exterior and is known as a gastrula. The opening of the gastrula is the blastopore, or primitive mouth; both terms are somewhat misleading. It would seem that the term blastopore should be applied more appropriately to an opening in a blastula, in which, of course, no opening exists. As to the term primitive

The gastrula stage

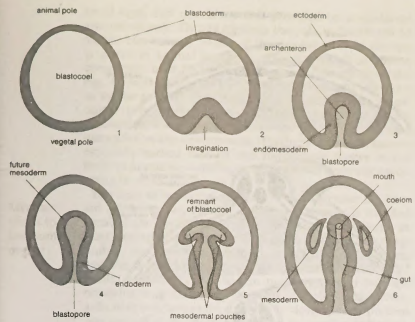


Figure 15: Gastrulation and mesoderm formation in the embryo of a starfish.

mouth, it must be pointed out that the blastopore does not always give rise to the adult mouth. In certain animal groups it becomes the anus, and a mouth forms as a completely new opening.

In some coelenterates, cells at the vegetal pole do not form an invaginating pocket, but individual cells slide inward, losing connection with other cells of the blastoderm. Eventually these cells fill the blastocoel and form a compact mass of endoderm. The cavity (archenteron) within this mass and the opening (blastopore) to the exterior are then produced secondarily by the separation of these cells.

Amphioxus, echinoderms, and amphibians. Gastrulation does not always proceed exactly as described above. In the course of evolution, certain animal groups have modified this critical stage of embryonic development, and these modifications have undoubtedly contributed to the successful continuation of species. In the primitive fish-like chordate amphioxus, for example, the invaginating blastoderm eventually comes into close contact with the inner surface of the ectoderm, thus practically squeezing the blastocoel out of existence or at least reducing it to a narrow crevice between the ectoderm and the endomesoderm. In echinoderms, on the other hand, a smaller portion of the blastoderm invaginates, and the blastocoel remains as a spacious internal cavity between the ectoderm and the endomesoderm. It persists as the primary body cavity and is the only body cavity (apart from the cavity of the alimentary canal) in such invertebrates as nematodes and rotifers.

In the double-walled-cup stage, the two internal germinal layers—endoderm and mesoderm—may not yet be distinct. Their separation may occur later, in the second phase of gastrulation, by one of two methods. One is the development of outpocketings from the wall of the archenteron. In starfishes and other echinoderms, the deep part of the endomesodermal invagination forms two thin-walled sacs, one on each side of the gastrula (Figure 15). These are the rudiments of the mesoderm; the remaining part of the archenteron becomes the endoderm and produces the lining of the gut. The cavities within the mesodermal sacs expand to become the coelom, the secondary body cavity of the animal. A somewhat similar process of mesoderm and coelom development occurs in amphioxus among the chordates, except that a series of mesodermal sacs forms on either side of the embryo, foreshadowing the segmented (metameric) structure common to chordates. Only the most anterior pairs of the mesodermal sacs actually contain a cavity at the time of their formation; the more posterior ones are solid masses of cells separating from the archenteric wall and from one another and developing coelomic cavities later.

Gastrulation in starfishes

A second method of mesoderm formation is by the splitting off of mesodermal cells from the original common mass of endomesoderm. This may take the form of single cells detaching themselves from the archenteron or of whole sheets of cells splitting off from the endoderm. An example of the latter type is seen in the gastrulation of amphibians. The development of specific regions of the early amphibian embryo—by the use of natural pigmentation or artificially introduced dyes—can be followed and their location in the adult recorded in diagrams called fate maps. The fate map of a frog blastula just prior to gastrulation demonstrates that the materials for the various organs of the embryo are not yet in the position corresponding to that in which the organs will lie in a fully developed animal. The endodermal material for the foregut, for example, lies not far from the vegetal pole; the ectodermal component of the mouth region (stomodeum) is situated close to the animal pole. Extensive rearrangement of the embryo is necessary to bring all the parts into their correct relationships.

Because of the large amount of yolk and resulting uneven cleavage, gastrulation in amphibians cannot proceed by a simple infolding of the vegetal hemisphere. A certain amount of invagination does take place, assisted by an active spreading of the animal hemisphere of the embryo; as a result, the ectoderm covers the endodermal and mesodermal areas. The spreading is sometimes described as an "overgrowth"—an inappropriate term, since no growth or increase of mass is involved. The future ectoderm simply thins out, expands, and covers a greater surface of the embryo in a movement known as epiboly.

Gastrulation in amphibians, in lungfishes, and in the cyclostomes (hagfishes and lampreys) begins with the formation of a pit on what will become the back (dorsal) side of the embryo (Figure 16). The pit represents the active

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970), W.B. Saunders Company, Philadelphia

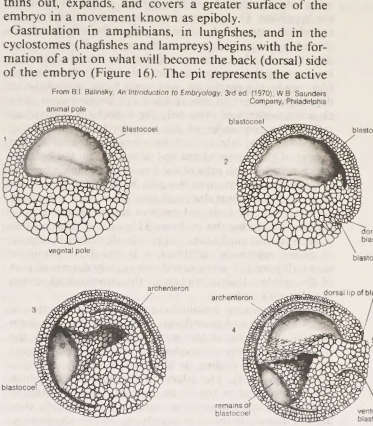


Figure 16: Gastrulation in the frog embryo. Embryos, in four progressive stages, are represented as cut in the median plane.

shifting inward of the cells of the blastoderm. As these cells undergo a change in shape, there occurs also a contraction at the external surface, with adjacent cells being drawn toward the centre of the contraction even before an actual depression is formed. The cells most concerned in this process will become part of the future foregut. Further movement of the cells inward results in the formation of a distinct pit, which rapidly develops into a pocket-like archenteron with its opening, the blastopore. Once the archenteron is formed, more and more of the exterior cells roll over the edge of the blastopore and disappear into the interior. In the course of gastrulation the shape of the blastopore changes from a simple pit to a transverse slit and finally into a groove encircling the yolky material at the vegetal pole. As a result of epiboly of the animal hemisphere, the upper edge of the groove is gradually pushed down until the yolky cells of the vegetal pole are covered completely. The edges of the blastopore then converge toward the vegetal pole, the slit between them

Formation of archenteron in amphibians

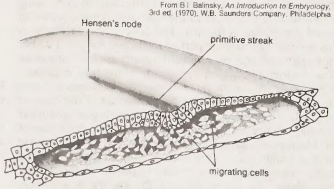
being eventually reduced to a narrow canal, which lies at the posterior end of the embryo and, in some species, becomes the anal opening. (In other cases the canal closes, and a new anal opening breaks through nearby, slightly more ventrally.)

The cavity of the archenteron increases as more material from the outside is transferred inward, and the blastocoel becomes almost completely obliterated. Both mesoderm and endoderm are shifted into the interior, and only the ectoderm remains on the embryo surface. The mesoderm splits from the endoderm: the endoderm lines the archenteric cavity (and eventually becomes the lining of the alimentary canal), as the mesoderm surrounds the endoderm to form the chordamesodermal mantle. By the time the blastopore closes, the three germ layers are in their correct spatial relationship to each other.

Reptiles, birds, and mammals. Although amphibian gastrulation is considerably modified in comparison with that in animals with oligolecithal eggs (*e.g.*, amphioxus and starfishes), an archenteron forms by a process of invagination. Such is not the case, however, in the higher vertebrates that possess eggs with enormous amounts of yolk, as do the reptiles, birds, and egg-laying mammals. Cleavage in these animals is partial (meroblastic), and, at its conclusion, the embryo consists of a disk-shaped group of cells lying on top of a mass of yolk. This cell group often splits into an upper layer, the epiblast, and a lower layer, the hypoblast. These layers do not represent ectoderm and endoderm, respectively, since almost all the cells that form the embryo are contained in the epiblast. Future mesodermal and endodermal cells sink down into the interior, leaving only the ectodermal material at the surface. In reptiles, egg-laying mammals, and some birds, a pocket-like depression occurs in the epiblast but encompasses only chordamesoderm or even only the notochord. Individual cells of the remainder of the mesoderm and endoderm migrate into the interior and there arrange themselves into a sheet of chordamesoderm and of endoderm, the latter of which mingles with cells of the hypoblast if such a layer is present. The migration of the cells destined to form mesoderm and endoderm does not take place over the whole surface of the disk-shaped embryo but is restricted to a specific area along the midline. This area is more or less oval in reptiles and lower mammals; distinctly elongated in higher mammals and birds, it is called the primitive streak (Figure 17), a thickened and slightly depressed part of the epiblast that is thickest at the anterior end, called the Hensen's node.

In animals having discoidal cleavage, the three germinal layers at the end of gastrulation are stacked flat; ectoderm on top, mesoderm in the middle, and endoderm at the bottom. The embryo is produced from the flattened layers by a process of folding to form a system of concentric tubes (Figure 18). The edges of the germ layers, which are not involved in the folding process, remain attached to the yolk and become the extra-embryonic parts; they are not directly involved in supplying cells for the embryo but break down yolk and transport it to the developing embryo.

Higher mammals—apart from the egg-laying mammals—do not have yolk in their eggs but, having passed through an evolutionary stage of animals with yolky eggs, retain,



From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970), W.B. Saunders Company, Philadelphia

Figure 17: Anterior half of blastoderm of a bird embryo cut transversely to show migration of mesodermal and endodermal cells from the primitive streak.

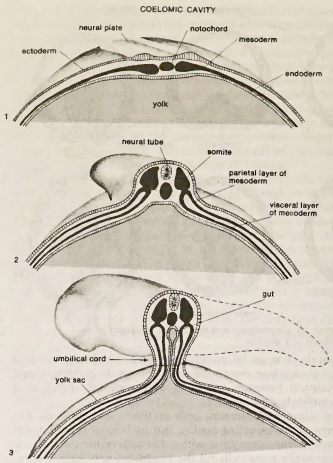


Figure 18: Vertebrate body formation from three germinal layers showing separation from yolk sac after discoidal cleavage.

particularly in gastrulation, features common to reptiles (and birds, which also had reptilian ancestors). As a result, at the end of cleavage the formative cells of the embryo—the cells that will actually build the body of the animal—are arranged in the form of a disk over a cavity that takes the place of the yolk of the reptilian ancestors of mammals. Within the disk of cells a primitive streak develops, and the three germinal layers are formed much as in many reptiles and birds.

Gastrulation in higher mammals

Gastrulation and the formation of the three germinal layers is the beginning of the subdivision of the mass of embryonic cells produced by cleavage. The cells then begin to change and diversify under the direction of the genes. The genes brought in by the sperm exert control for the first time; during cleavage all processes seem to be under control of the maternal genes. In cases of hybridization, in which individuals from different species produce offspring, the influence of the sperm is first apparent at gastrulation: paternal characteristics may appear at this stage; or the embryo may stop developing and die if the paternal genes are incompatible with the egg (as is the case in hybridization between species distantly related).

The diversification of cells in the embryo progresses rapidly during and after gastrulation. The visible effect is that the germinal layers become further subdivided into aggregations of cells that assume the rudimentary form of various organs and organ systems of the embryo. Thus the period of gastrulation is followed by the period of organ formation, or organogenesis.

EMBRYONIC ADAPTATIONS

Throughout its development the embryo requires a steady supply of nourishment and oxygen and a means for disposal of wastes. These needs are met in various ways, depending in particular on (1) whether the eggs develop externally (oviparity), are retained in the maternal body until ready to hatch (ovoviviparity), or are carried in the maternal body to a later stage (viviparity); and (2) the length of embryonic development.

Adaptations in animals other than mammals. Eggs of many marine invertebrates are discharged directly into water, and the period of development before the larva emerges is relatively brief. Oxygen diffuses easily into the

small eggs, and nourishment is provided by a moderate amount of yolk. During cleavage the yolk is distributed to all the blastomeres. Much of the nourishment in the egg is stored as animal starch, or glycogen, which is almost completely used by the time the larva emerges from the egg. A small amount of water and inorganic salts are taken in by the embryo from surrounding seawater. Eggs developing in freshwater carry their own supply of necessary amounts of certain salts that are not present in sufficient quantities in the environment. Products of metabolism—especially carbon dioxide and nitrogenous wastes in the form of ammonia—diffuse out from small embryos developing in water.

Mechanisms to overcome drying

The eggs of terrestrial animals must overcome the hazard of drying. In certain species this danger is avoided because the animal returns to water to breed, such as frogs and salamanders. Some groups of insects (*e.g.*, dragonflies, mayflies, and mosquitoes) also lay eggs in water, and the larvae are aquatic. Eggs of other animals (*e.g.*, snails, earthworms) are laid in moist earth and thus are protected from drying up. In terms of evolution, however, a decisive solution to the problem of development on land was arrived at by most insects and by reptiles and birds, which developed eggs with a shell impermeable to water or, at least, resistant to rapid evaporation. The shells of bird and insect eggs, while restricting evaporation of water, allow oxygen to diffuse into the egg and carbon dioxide to diffuse out. Apart from gas exchange, the eggs constitute closed systems, which give nothing to the outside and require nothing from it. Such eggs are called cleidoic. Because the products of nitrogen metabolism in cleidoic eggs cannot pass through the eggshell, animals (birds and insects) have had to evolve a method of storing wastes in the form of uric acid, which, since it is insoluble, is nontoxic to the embryo.

After a short period of development in the egg, the emerging young animal has to fend for itself, unless there is some form of parental care. Exposure to the external environment at a tender age results frequently in loss of life, a hazard met by many animals through an increase in the supply of nourishment within the egg, thus allowing the young to attain a greater size and development. This tendency to produce large yolky eggs has been achieved independently in different evolutionary lines: in octopuses and squids among the mollusks, in sharks among the fishes, and in reptiles and birds among the terrestrial vertebrates.

As has been indicated, cleavage is incomplete in eggs with large amounts of yolk. Although some yolk platelets may be enclosed in the formative cells of the embryo, the bulk of the yolk remains an uncleaved mass, overgrown and surrounded by the cellular part of the embryo. In such cases a membranous bag, or yolk sac, is formed (Figure 19) and remains connected to the embryo by a narrow stalk (the evolutionary precursor of the umbilical cord of mammals). The cellular layers surrounding the yolk sac and forming its walls may consist of all three germinal layers (in reptiles and birds), so that the yolk virtually comes to lie inside an extension of the gut of the embryo; or (in bony fishes) the yolk sac may be enclosed in layers of ectoderm and mesoderm. In either case a network of blood vessels develops in the walls of the yolk sac and transports the yolk products to the embryo. As the yolk is broken down and utilized, the yolk sac shrinks and is eventually drawn into the body of the embryo. In addition to the yolk sac, extra-embryonic parts are also encountered in the form of embryonic membranes, which are found in higher vertebrates and in insects. Vertebrates have three embryonic membranes: the amnion, the chorion, and the allantois.

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970), W.B. Saunders Company, Philadelphia

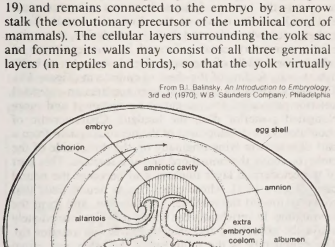


Figure 19: Position of a bird embryo in the egg and the relation of the allantois to the amniotic cavity and the yolk sac.

In reptiles, birds, and mammals, folds develop on the surface of the yolk sac just outside and around the body of the embryo proper. These folds, consisting of extra-embryonic ectoderm and extra-embryonic mesoderm, rise up and fuse dorsally, enclosing the embryo in a double-lined, fluid-filled chamber known as the amniotic cavity. The inner lining of the fold becomes the amnion, and the outer becomes the chorion, which ultimately surrounds the entire embryo. The amniotic fluid protects the embryo from drying, prevents the adhesion of the embryo to the inner surface of the shell, and provides the embryo with efficient shock absorption against possible damaging jolts. (The aminion and chorion develop in the same way in insect embryos.) The third membrane, or allantois, is originally nothing more than the urinary bladder of the embryo. It is a saclike growth of the floor of the gut, into which nitrogenous wastes of the embryo are voided. It enlarges greatly during the course of development, eventually expanding between the amnion and chorion and also between the chorion and the yolk sac, to become the third embryonic membrane. In addition to providing storage space for the nitrogenous wastes of the embryo, the allantois takes up oxygen, which penetrates into the egg from the exterior, and delivers it, by way of a network of blood vessels, to the embryo.

The allantois

Adaptations in mammals. At some early stage during the evolution of viviparous mammals, eggs came to be retained in the oviducts of the mother. The embryo then was provided with nourishment from fluids in the oviduct; the yolk, which became redundant, gradually ceased to be provided, and the eggs became oligolecithal. The eggshell, present in reptiles, was no longer needed and eventually disappeared, as did the white of the egg. The chorion, however, remained as the most external coat of the developing embryo through which nourishment reaches the embryo. It acquired the ability to adhere closely to the walls of the uterus (which was what that part of the oviduct holding the embryo had become) and became the so-called trophoblast. The blood-vessel network of the underlying allantois conveys nutrients that diffuse through the trophoblast to the body of the embryo proper. These modifications gave rise to a new organ, the placenta, formed from tissues of both the mother and the embryo: the uterine wall with its blood vessels provided by the mother; the trophoblast and allantois—and in some mammals also the yolk sac—with their blood vessels provided by the embryo.

The overall development of placental mammals as a result of these changes is profoundly different from that of their ancestors, the reptiles, and proceeds in the following way: the tiny yolkless egg is fertilized in the upper portion of the oviduct by sperm received from the male in the process of coupling (coitus); cleavage starts as the egg is propelled slowly down the oviduct by action of cilia in the oviduct lining. At the end of cleavage a solid ball of cells called a morula is produced. The surface cells of the morula become the trophoblast and the inner cell mass gives rise to the embryo (the formative cells) and also its yolk sac, amnion, and allantois. A cavity appears within the morula, converting it into a hollow embryo, called the blastocyst. This cavity resembles the blastocoel but, in fact, is analogous to the yolk sac of meroblastic eggs, except that there is no yolk and the cavity is filled with fluid. At the blastocyst stage, the embryo enters the uterus and attaches itself to the uterine wall. This attachment, or implantation, a crucial step in the development of a mammal, is attained through the action of the trophoblast,

Implantation

which forms extensions, known as villi, that penetrate the uterine wall (Figure 20). In higher placental mammals, the lining of the uterine wall and, in varying degrees, the underlying tissues as well are partially destroyed, resulting in a closer relationship between the blood supplies of the mother and the embryo. Indeed, in man and in some rodents, the blastocyst sinks completely into the uterine wall and becomes surrounded by uterine tissue.

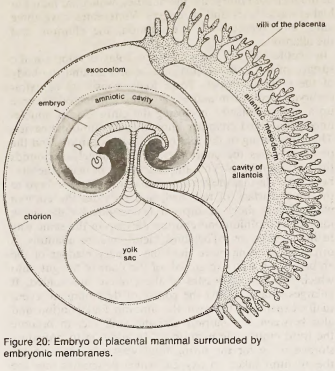


Figure 20: Embryo of placental mammal surrounded by embryonic membranes.

While implantation takes place, the formative cells arrange themselves in the form of a disk under the trophoblast. In the disk, the germinal layers develop much as in birds, with the formation of a primitive streak and migration of the chordamesoderm into a deeper layer. A layer of endoderm is formed adjoining the cavity of the blastocyst, and an amniotic cavity develops, enclosing the embryo; in lower placental mammals, the allantois also develops. The embryo proper, lying in the amniotic cavity, is connected to the extra-embryonic parts by the umbilical cord. The umbilical cord lengthens greatly during later development. In higher mammals, the cavity of the allantois is reduced, but the allantoic blood vessels become well developed and extend through the umbilical cord, connecting the embryo to the placenta. The blood that circulates in the placenta brings oxygen and nutrients from the maternal blood to the embryo and carries away carbon dioxide and other waste products from the embryo to the maternal blood for disposal by the maternal body.

Although tissues of maternal and embryonic origin are closely apposed in the placenta, there is little actual mingling of the tissues. Despite an occasional penetration of an embryo cell into the mother and vice versa, there is a placental barrier between the two tissues. The blood circulation of the mother is at all times completely separated from that of the embryo and its extra-embryonic parts. The placental barrier, however, does allow molecules of various substances to pass through; such differential permeability is indeed necessary if the embryo is to obtain nourishment. The permeability of the placental barrier differs in different animals; thus antibodies, which are protein molecules, may penetrate the placental barrier in man but not in cattle.

The maintenance of the fetus—as the more advanced embryo of a mammal is called—in the uterus is under hormonal control. In the initial stages of pregnancy, the continued existence of the embryo in the uterus depends on the hormone progesterone, which is secreted by the corpora lutea, "yellow bodies," that develop in the ovary after an egg has been released.

At birth the fetal parts of the placenta separate from the maternal parts. Contraction of the uterine wall first releases the fetus from the uterus; the fetal parts of the placenta (the afterbirth) follow. In certain cases of intimate connection between fetal and maternal tissues, the maternal tissues are torn, and birth is accompanied by profuse bleeding.

Organ formation

PRIMARY ORGAN RUDIMENTS

Immediately after gastrulation—and sometimes even while gastrulation is underway—the germinal layers begin subdividing into regions that will give rise to various parts of the body. Subdivision proceeds in stages: initially a mass of cells is set aside for an organ system (for the alimentary canal, for instance) and subsequently further subdivided into the rudiments of various parts of the organ system, such as the liver, stomach, and intestines. The initially formed larger units are referred to as primary organ rudiments; those they later give rise to, as secondary organ rudiments.

The type of organ rudiment produced depends on the organization of the body in any particular group in the animal kingdom. In the vertebrates the earliest subdivision within a germinal layer is the segregation within the chordamesodermal mantle of the rudiment of the notochord from the rest of the mesoderm (Figure 18). During gastrulation the material of the notochord comes to lie middorsally in the roof of the archenteron. It separates by longitudinal crevices from the chordamesodermal mantle lying to the left and right. The material of the notochord then rounds off and becomes a rod-shaped strand of cells immediately under the dorsal ectoderm, stretching from the blastopore toward the anterior end of the embryo, to the midbrain level. In front of the tip of the notochord, there remains a thin sheet of prechordal mesoderm.

Development of notochord

The mesodermal layer adjoining the notochord becomes thickened and, by transverse crevices, subdivided into sections called somites (Figure 21). The somites, which later give rise to the segmented body muscles and the vertebral column, are the basis of the segmented organization typical of vertebrates (seen especially in the lower fishlike forms but also in the embryos of higher vertebrates). The lateral and ventral mesoderm, which remains unsegmented, is called the lateral plate. The somites remain connected to the lateral plate by stalks of somites that play a particular role in the development of the excretory (nephric) system in vertebrates; for this reason they are called nephrotomes. Rather early the mesodermal mantle splits into two layers, the outer parietal (somatic) layer and the inner visceral (splanchnic) layer, separated by a narrow cavity (Figure 22) that will expand later to form the coelomic, or secondary, body cavity. The coelomic cavity extends initially through the nephrotomes into the somites; in the somites it is eventually obliterated. Endoderm completely surrounds the lumen of the archenteron (when present) and produces the cavity of the alimentary canal. If no archenteric cavity is formed during gastrulation, the cavity of the alimentary canal is formed by the separation of cells in the middle of the mass of endoderm (as in bony fishes) or by folding of the sheet of endoderm (Figure 18). The endodermal gut sooner or later acquires an extended anterior part called the foregut and a narrower and more elongated posterior part, the hindgut. Characteristic of chordates is the development of the nervous system from a part of ectoderm lying originally on the dorsal side of the embryo, above the notochord and the somites. This part of the ectodermal layer thickens and becomes the neural plate (Figure 18:1), whose edges rise as neural folds that converge toward the midline, fuse together, and form the neural tube. In vertebrates the neural tube lies immediately above the notochord and extends beyond its anterior tip. The neural tube is the rudiment of the brain and spinal cord; its lumen gives rise to the cavities, or ventricles, of the brain and to the central canal of the spinal cord. The remainder of the ectoderm closes over the neural tube and becomes, in the main, the covering layer (epithelium) of the animal's skin (epidermis). As the neural tube detaches itself from the overlying ectoderm, groups of cells pinch off and form the neural crest, which plays an important role in the development of, among other things, the segmental nerves of the brain and spinal cord.

Formation of the neural crest

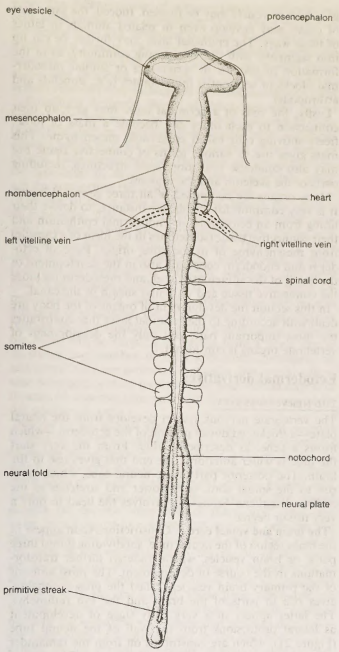


Figure 21: Early chick embryo, showing development of the brain vesicles and of some of the mesodermal structures.
From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970). W.B. Saunders Company, Philadelphia

In developing the primary organ rudiments mentioned above, the embryo acquires a definite organization clearly recognizable as that of a chordate animal. Similar processes, which occur in the development of other animals, establish the basic organization of an annelid, a mollusk, or an arthropod.

The organization of the embryo as a whole appears to be determined to a large extent during gastrulation, by which process different regions of the blastoderm are displaced and brought into new spatial relationships to each other. Groups of cells that were distant from each other in the blastula come into close contact, which increases possibilities for interaction between materials of different origin. In the development of vertebrates in particular, the sliding of cells (presumptive mesoderm) into the interior and their placement on the dorsal side of the archenteron (in the archenteric "roof"), in immediate contact with the overlying ectoderm, is of major importance in development and subsequent differentiation. Experiments have shown that, at the start of gastrulation, ectoderm is incapable of progressive development of any kind; that only after invagination, with chordamesoderm lying directly underneath it, does ectoderm acquire the ability for progressive development. The dorsal mesoderm, which later differentiates into notochord, prechordal mesoderm, and somites, causes the overlying ectoderm to differentiate as neural plate. Lateral mesoderm causes overlying ectoderm to differentiate as skin. The influence exercised by parts of the embryo, which causes groups of cells to proceed along a particular path of development, is called embryonic induction. Though induction requires that the interacting parts come into close proximity, actual contact is not necessary. The inducing influence—whatever it might be—is a diffusible substance emitted by the activating cells (the inductor). The inducing substance of the mesoderm is a large molecule, probably a protein or a nucleoprotein, which presumably penetrates reacting cells, though direct and unequivocal proof of such penetration is still unavailable. Inducing substances are active on vertebrates belonging to many different classes; *e.g.*, inductions of primary organs have been obtained by transplanting mammalian tissues into frog embryos or by transplanting tissues of a chick embryo into the embryo of a rabbit.

Induction is responsible not only for the subdivision of ectoderm into neural plate and epidermis but also for the development of a large number of organ rudiments in vertebrates. The notochord is a source of induction for the development of the adjoining somites and nephrotomes; the latter appear jointly to induce development of limb rudiments from the lateral plate mesoderm. Further examples are mentioned below in connection with development of the various organs.

Inducing substances

Since the results of induction are different for different organ rudiments, it must be presumed that there exist inducing substances with specific action, at least to a certain extent; thus, the lateral mesoderm induces differentiation of the skin but not neural plate from the very same kind of ectoderm. The number of inducing substances need not, however, be the same as the number of different kinds of tissues and organs, since certain differentiations could possibly be induced by a combination of two or more inducing substances, or the same inducing substance might have different effects on different tissues. It has been suggested that the regional organization of the entire vertebrate body could be controlled by the graded distribution of only two inducing substances—provisionally named the neuralizing substance and the mesodermalizing substance—along the length of the embryo. The neuralizing substance, concentrated at the anterior end, gradually decreases toward the posterior end; the mesodermalizing substance, on the

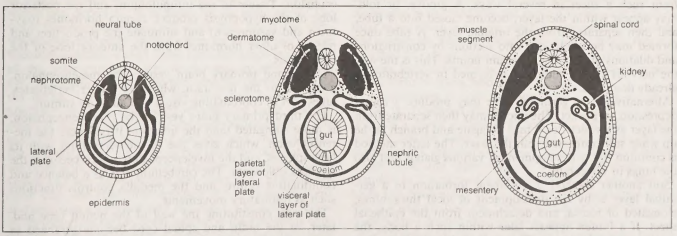


Figure 22: (Left to right) Subdivision of the mesoderm in a vertebrate embryo.

other hand, is concentrated at the posterior end and decreases toward the anterior end. The differentiation of induced structures depends on the relative amounts of the two inducing substances at any given point in the embryo. Acting alone, the neuralizing substance induces only nervous tissue, which takes the form of the forebrain, and the mesodermalizing substance induces only mesodermal structures (*e.g.*, somites, notochord).

In the amphibian embryo, induction appears to have its primary source in the dorsal lip of the blastopore, which eventually gives rise to the notochord and adjoining somites. Induction by the notochord and somites is responsible for the development of the neural plate in the ectoderm, of lateral and ventral parts of the mesodermal mantle, and of the lumen of the alimentary canal in the endoderm. The dorsal lip of the blastopore for this reason has been called the primary organizer. In higher vertebrates, in which gastrulation occurs through the medium of a primitive streak, the anterior end of the streak and the Hensen's node have properties similar to those of a primary organizer. Organization centres have been found, or suspected, in embryos of animals belonging to a few other groups, in particular the insects and sea urchins, but the interpretation of the experimental results in these animals is less satisfactory than in the case of vertebrates.

Concept of organization centre

The concept of an organization centre suggests that a part of the embryo differs from the rest of the embryonic tissues in being more active. The more active parts of the embryo (and also of animals in later stages of development) are particularly sensitive to certain noxious influences in their environment. If an embryo is deprived of oxygen or subjected to weak concentrations of poisons, the first parts to suffer are the most morphogenetically active ones. In vertebrate embryos the anterior end of the head is most sensitive. Early sea-urchin embryos have two centres of maximal sensitivity: one at the animal pole and the other at the vegetal pole. The damage done by noxious influences may result in actual breakdown of cells in a region of maximal sensitivity and may also lead to a depression of the developmental potential of the cells. Thus, the graded distribution of certain physiological properties appears to play a part in morphogenetic processes: physiological gradients are in fact also morphogenetic gradients.

Gradients in the embryo can be used to control development to a certain extent, by exposing the embryo to influences that, while reaching all parts, have a local effect as the result of differences in sensitivity. Disturbances of normal development often are the result of disruptions of gradients (see above *Aberrations of growth: biological malformation.*).

ORGANOGENESIS AND HISTOGENESIS

The primary organ rudiments continue to give rise to the rudiments of the various organs of the fully developed animal in a process called organogenesis. The formation of organs, even those of diverse function, shares some common features, which are considered in this section. As the organs form, so do their component tissues, in a process termed histogenesis.

A germinal layer, as the name implies, is a sheet of cells. An organ rudiment may be formed and separated from such a sheet in several ways. A groove, or fold, may appear within the layer, become closed into a tube, and then separated from the original layer. A tube once formed may be subdivided into sections by constrictions and dilations of the tube at certain points. This is the way the nervous system rudiment is formed in vertebrates as already described.

Alternatively, the germinal layer may produce a round depression, or pocket. The pocket may then separate from the layer as a vesicle, or it may elongate and branch at the tip while still connected with the layer. The latter method is common in the development of various glands and also the lungs in vertebrates.

Still another method of rudiment formation in a germinal layer is by the development of local thickenings, elongated or round, and detachment from the epithelial sheet. If a lumen appears later within such a body, the result may be the same as that achieved by folding—that is, a tube or vesicle may be formed. Indeed, the same sort of organ may develop even in related animals in either of these ways. The epithelial layer may further be cut up into segments, with the layer losing continuity, as in the formation of somites in vertebrates or similar mesodermal blocks in segmented invertebrates (*e.g.*, annelids and arthropods).

Embryonic mesenchyme

Lastly, the cells of a germinal layer may give up their connection to each other and become a mass of loose, freely moving cells called embryonic mesenchyme. This mass gives rise to various forms of connective tissue but may also condense into more solid structures, including parts of the skeleton and the muscles.

Many organs are comprised of all three germinal layers. It is very common for glands, for instance, to derive their lining from an ectodermal or endodermal epithelium and their connective tissue (sometimes in the form of a capsule) from mesenchyme of mesodermal origin. Parts of ectoderm and endoderm cooperate also in the development of the lining of the alimentary canal, and mesoderm provides the connective tissue and muscular sheath of the canal.

In this section the development of organs of the body are dealt with according to the germinal layer that contributes the most important part, and only the development of vertebrate organs is considered.

Ectodermal derivatives

THE NERVOUS SYSTEM

The vertebrate nervous system develops from the neural plate—a thickened dorsal portion of the ectoderm—which forms a tube, as described earlier. From the very start the tube is wider anteriorly, the end that gives rise to the brain. The posterior part of the neural tube, which gives rise to the spinal cord, is narrower and stretches as the embryo lengthens. Stretching involves the head to only a very minor degree.

The brain and spinal cord. Constrictions soon appear in the brain region of the neural tube, subdividing it into three parts, or brain vesicles, which undergo further transformations in the course of development. The most anterior of the primary brain vesicles, called the prosencephalon, gives rise to parts of the brain and the eye rudiments. The latter appear in a very early stage of development as lateral protrusions from the wall of the neural tube (Figure 21), which are constricted off from the remainder of the brain rudiment as the optic vesicles. The rest of the prosencephalon constricts further into two portions, an anterior one, or telencephalon, and a posterior one, or diencephalon. The telencephalon gives rise, in lower vertebrates, to the smell, or olfactory, centre; in higher vertebrates and man, it becomes the centre of mental activities. The diencephalon, with which the eye vesicles are connected, was presumably originally an optic centre, but it has acquired, in the course of evolution, a function of hormonal regulation. The floor of the diencephalon forms a funnel-shaped depression, the infundibulum, which becomes connected with the pituitary, or hypophysis, the most important gland of internal secretion (*i.e.*, endocrine gland) in vertebrates. Indeed, the posterior lobe of the hypophysis is actually derived from the floor of the diencephalon. Tissues of the infundibulum and the posterior lobe of the hypophysis produce certain hormones (oxytocin and vasopressin) and stimulate the production and release of other hormones from the anterior lobe of the hypophysis.

The second primary brain vesicle, the mesencephalon, gives rise to the midbrain, which, in higher vertebrates, takes part in coordinating visual and auditory stimuli.

The third primary brain vesicle, the rhombencephalon, is more elongated than the first two; it produces the metencephalon, which gives rise to the cerebellum with its hemispheres, and the myelencephalon, which becomes the medulla oblongata. The cerebellum acts as a balance and coordinating centre, and the medulla controls functions such as respiratory movements.

The cells constituting the wall of the neural tube and, later, of the brain and spinal cord become arranged in such a way that they point into the central cavity of the

Differentiation of nervous tissue

tube. The differentiation of nervous tissue involves many cells abandoning their connection to the inner surface of the neural tube and migrating outward, where they accumulate as a mantle. The first cells to migrate become the neurons, or nerve cells. They produce outgrowths called axons and dendrites, by which the cells of the nervous system establish communication with one another to form a functional network. Some of the outgrowths extend beyond the confines of the brain and spinal cord as components of nerves; they establish contact with peripheral organs, which thus fall under the control of the nervous system. Cells migrating from the inner surface of the neural tube later in development become astrocytes, which are the supporting elements of nerve tissue.

The fate of nerve cells is dependent largely on whether they succeed, directly or indirectly (through other neurons), in connecting with peripheral organs. Nerve cells that fail to establish connections die. Thus, if in early stages of embryonic development, some organ, a limb rudiment for instance, is surgically removed, the nerve cells in the centres supplying nerves to such an organ are reduced in number, and the corresponding nerves also diminish or disappear. On the other hand, if an organ is introduced by transplantation into a developing embryo, the organ will be supplied by nerves from a nerve centre in which the number of cells apparently increases: no additional cells are provided, but cells that would otherwise have degenerated remain active and differentiate into functional neurons, thus satisfying the demand created by the additional organ.

Nerves do not consist entirely of outgrowths of neurons located in the brain and spinal cord. Many components of nerves are outgrowths of neurons, the cell bodies of which are located in masses called ganglia; there are three main types of ganglia: spinal ganglia, cranial ganglia, and ganglia of the autonomous nervous system. The spinal ganglia are derived from cells of the neural crest—the loose mesenchyme-like tissue that remains between the neural tube and skin after separation of the two. Part of the cells of the neural crest in the region of the trunk and tail accumulate in segmental groups (corresponding to the mesodermal somites) and provide fibres to peripheral organs and to the spinal cord. These fibres constitute the sensory pathways in the spinal nerves. The motor components of the spinal nerves—fibres that activate muscles—are outgrowths of neurons lying in the spinal cord. The ganglia of the cranial nerves are produced only in part from cells of the neural crest; an additional component comes from the epidermis on the side of the head. Cells of the epidermal thickenings called placodes detach themselves and contribute to the formation of the cranial ganglia and thus of the cranial nerves.

The ganglia of the autonomous (sympathetic) nervous system are derived, as are the spinal ganglia, from neural-crest cells, but, in this case, the cells migrate downward to form groups near the dorsal aorta, near the intestine, and even in the intestinal wall itself. The outgrowths of cells in these ganglia are the nerve fibres of the sympathetic nerves (see also NERVES AND NERVOUS SYSTEMS).

Major sense organs. *The eye.* As has been pointed out, the rudiments of the eyes develop from optic vesicles, each of which remains connected to the brain by an eye stalk, which later serves as the pathway for the optic nerve. The optic vesicles extend laterally until they reach the skin, whereupon the outer surface caves in so that the vesicle becomes a double-walled optic cup. The thick inner layer of the optic cup gives rise to the sensory retina of the eye; the thinner outer layer becomes the pigment coat of the retina. The opening of the optic cup, wide at first, gradually becomes constricted to form the pupil, and the edges of the cup surrounding the pupil differentiate as the iris. The refractive system of the eye and, in particular, the lens of the eye are derived not from the cup but from the epidermis overlying the eye rudiment. When the optic vesicle touches the epidermis and caves in to produce the optic cup, the epidermis opposite the opening thickens and produces a spherical lens rudiment. The lens develops by an induction by the optic vesicle on the epidermis with which it comes in contact. A further influence emanating from the eye changes the epidermis remaining in place over the lens into a transparent area, the cornea. Influence of the optic cup on the surrounding mesenchyme causes the latter to produce a vascular layer around the retina and, outside of that, a tough fibrous or (in some animals) even a partly bony capsule called the sclera. Thus a complex interdependence of different materials produces the fully developed and functional vertebrate eye (see also SENSORY RECEPTION: *Photoreception*).

Formation of the optic cup

The ear. The main part of the ear rudiment is derived from thickened epidermis adjoining the medulla. This area of the epidermis invaginates to produce the ear vesicle, which separates from the epidermis but remains closely apposed to the medulla. The ear vesicle becomes complexly folded to produce the labyrinth of the ear. Subsequently, a group of cells of the ear vesicle becomes detached and gives rise to the acoustic ganglion. Neurons of this ganglion become connected by their nerve fibres to the sensory cells in the labyrinth, on the one hand, and with the brain (the medulla), on the other. The ear vesicle, acting on the surrounding mesenchyme, induces the latter to aggregate around the labyrinth and form the ear capsule. Further parts with various origins are added to the ear: the middle ear, from a pharyngeal pouch and the associated skeleton, and the external ear (where present), from epidermis and dermis.

The olfactory organ. The olfactory organ develops from a thickening of the epidermis adjacent to the neural fold at the anterior end of the neural plate. This thickening is converted into a pocket or sac but does not lose connection with the exterior. The openings of the sac become the external nares, and the cavity of the sac becomes the nasal cavity. Some cells of the olfactory sac differentiate as sensory epithelium and produce nerve fibres entering the forebrain. In most fishes the olfactory sac does not communicate with the oral cavity; in lungfishes and in terrestrial vertebrates, however, canals develop from the olfactory sacs to the oral cavity, where they open by internal nares. A cartilaginous capsule forms around the olfactory organ from cells believed to have been derived from the walls of the sac itself, and thus it is ectodermal in origin.

The nares and nasal cavity

Gustatory and other organs. Gustatory organs in the form of taste buds develop as local differentiations of the lining of the oral cavity but also, in fishes, in the skin epidermis. They are supplied with nerve endings, as are several other sensory bodies scattered among the tissues and organs of the developing body.

THE EPIDERMIS AND ITS OUTGROWTHS

The major part of the ectodermal epithelium covering the body gives rise to the epidermis of the skin. In fishes and aquatic larvae of amphibians, the many-layered epidermis is provided with unicellular mucous glands. In terrestrial vertebrates, however, the epidermis becomes keratinized; *i.e.*, the outer layers of cells produce keratin, a protein that is hardened and is impermeable to water. During the process of keratinization, many cell components degenerate and the cells die; the layer of keratinized cells is therefore shed from time to time. In reptiles the shedding may take the form of a molt in which the animal literally crawls out of its own skin. It is less well known that frogs and toads also molt, shedding the surface keratinized layer of their skin (which is usually eaten by the animal). In birds and mammals, keratinized cells are shed in pieces that are sloughed off, rather than in extensive layers. In many vertebrates local thickenings of the keratinized layer appear in the form of claws, hooves, nails, and horns.

Keratinization

The epidermis is only the superficial layer of the skin, which is reinforced by the dermis, a connective tissue layer of a much greater thickness. The cells of the dermis are derived from mesoderm and neural-crest cells. In particular the pigment cells found in the dermis of fishes, amphibians, and reptiles are of neural-crest origin. The pigment in the skin of birds and mammals (and also in hairs and feathers) is also produced by neural-crest cells, but in these animals the pigment cells penetrate into the epidermis or deposit their pigment granules there.

The structure of the skin is further complicated by the

development of hairs and feathers, on the one hand, and of skin glands, on the other. Hairs and feathers develop from a somewhat similar kind of rudiment. The development starts with a local thickening of the epidermal layer, beneath which a group of mesenchyme cells accumulate. In the case of hairs, the epidermal thickening proliferates downward and forms the root of the hair, from which the shaft then grows outward, emerging on the surface of the skin. In the case of feathers, the epidermal thickening bulges outward to form a hollow fingerlike protrusion with a connective tissue core. Secondarily, the shaft of the feather branches characteristically to produce barbs and barbules. In both cases, however, the final structure—shaft of the hair and shaft barbs and barbules of the feather—consists of keratinized and, thus, dead cells.

The skin of amphibians and mammals (but not of birds and reptiles) is provided with numerous skin glands, which develop as ingrowths from the epidermis. A peculiar type of skin gland is the mammary gland of placental mammals. In the first stage of development, mammary-gland rudiments resemble hair rudiments; they are thickenings of the epidermis, with condensed mesenchyme on their inner surfaces. In some mammals (rabbit, man) two continuous epidermal thickenings called mammary lines stretch along either side of the belly of the embryo. Parts of the line corresponding in number and position to the future glands enlarge while the rest of the thickening disappears. The initial thickenings proliferate inward and produce a system of ramified cords, solid at first but hollowed out later, which become the lactiferous, or milk-bearing, ducts of the gland. Further branching at the tips of the ducts gives rise to smaller ducts and to the secretory end sacs, or alveoli, of the gland.

Mesodermal derivatives

THE BODY MUSCLES AND AXIAL SKELETON

Derivatives of somites

The somites, formed in the early stages of development from the upper edges of the mesodermal mantle adjoining the notochord, are complex rudiments that subdivide and give rise to very diverse body structures. The coelomic cavity, present initially, becomes obliterated by the side-to-side flattening of the somites, so that the thinner, outer parietal layer of the somite comes in close contact with its thicker visceral layer. The visceral layer of the somite very early subdivides into two parts. The upper, dorsolateral part called the myotome (Figure 22) remains compact, giving rise to the body muscles. The lower, medioventral part of the somite, called the sclerotome, breaks up into mesenchyme, which contributes to the axial skeleton of the embryo—that is, the vertebral column, ribs, and much of the skull. The parietal layer of the somite, at a later stage, is converted into mesenchyme that, together with components of the neural crest, gives rise to the dermis of the skin and, for this reason, is called the dermatome.

The cells of the myotome are elongated in a longitudinal direction and become differentiated as muscle fibres. The myotomes, originally situated dorsally, expand on either side, penetrating between the skin on the outside and the lateral plates of the mesoderm on the inside, until they meet midventrally; the whole body is thus enclosed in a layer of developing muscle. As the somites and myotomes are segmented, so are the muscles derived from them. Metamerism, or segmentation, a feature in the embryos of all vertebrates, remains preserved only in the adults of fishes and of terrestrial vertebrates that have elongated bodies (salamanders, snakes); it becomes largely erased in four-footed animals that depend on their limbs for locomotion.

The mesenchyme derived from the sclerotomes condenses as cartilage around the notochord and the spinal cord. It forms the cartilaginous vertebral column and ribs. In the head region it produces a part of the cartilaginous skull, mainly its posterior and ventral parts; anteriorly the somitic mesenchyme is supplemented by mesenchyme from the neural crest. Cartilaginous capsules of the olfactory organ and the ear fuse with the cartilaginous capsule surrounding the brain; to this complex are also added cartilages associated with the jaws and gill skeleton. Cartilage in the vertebral column and in the skull is replaced later in the bony fishes and in the terrestrial vertebrates by bone. At a still later stage, dermal bones are added, which, while they have no precursors in the cartilaginous skeleton, develop in the adjoining mesenchyme.

THE APPENDAGES: TAIL AND LIMBS

The tail in vertebrates is a prolongation of the body beyond the anus. It develops in early stages from the tail bud, immediately dorsal to the blastopore. Material for the tip of the tail is situated slightly forward from the edge of the blastopore. The elongation of the back of the body is greater than that of the belly; as a result the tip of the tail bud is carried beyond the blastopore and thus beyond the anus, which, in the developed embryo, marks the position of the blastopore. The consequence is that a section of the dorsal surface of the embryo comes to lie on the ventral surface of the tail; *i.e.*, becomes inflected. The tail bud is formed from parts that have already been differentiated to a certain extent; prolongations of the neural tube and of the notochord are involved, and endoderm extends into the tail rudiment as the postanal gut, which, however, soon degenerates. The bud is also encased in ectodermal epidermis. In amphibians the somites of the tail are not derived from the chordamesodermal mantle but from the inflected posterior portion of the neural plate, which loses its nervous nature and becomes subdivided into segments corresponding to the somites of the trunk. In higher vertebrates the cells in the interior of the tail bud have an undifferentiated appearance and form a growth zone, at the expense of which parts of the tail (neural tube, notochord, somites) are extended backward as the tail elongates.

Formation of the tail bud

The paired limbs of vertebrates derive their first rudiments from the upper edge of the lateral plate mesoderm. The parietal layer becomes thickened, and cells escape from the epithelial arrangement and form a mesenchymal mass adjoining the ectodermal epithelium at the surface of the body. The ectodermal epithelium over the mass of mesenchyme likewise becomes thickened. In higher vertebrates, the accumulation of mesodermal cells and the thickening of the epidermis occur along the entire length of the trunk, from neck to anus, but in the middle of the trunk they soon disappear, and only the most anterior and the most posterior sections develop further into the rudiments of the forelimbs and hindlimbs, respectively. In fishes, the rudiments of the pectoral and pelvic fins are more extended anteroposteriorly in earlier than in final stages.

The mesodermal masses of the limb rudiments proliferate, and, covered with thickened epidermis, form on the surface of the body conical protrusions called the limb buds, which, once formed, possess all the materials necessary for limb development. Limb buds may be transplanted into various positions on the body or on the head and there develop into clearly recognizable limbs, conforming to their origin, whether a forelimb or hindlimb, a wing or a leg in birds. This specificity of the limb is carried by the mesodermal part of the rudiment, but a complex interaction between the mesodermal mesenchyme and the ectodermal epidermis is necessary for the normal development of the limb. In four-limbed vertebrates (tetrapods), the tips of the limb buds become flattened and broadened into hand or foot plates. The edge of the plate is indented, forming the rudiments of the digits. Meanwhile, local areas of the mesodermal mesenchyme in the interior of the limb rudiment condense; these are the rudiments of the various components of the limb skeleton. In fishes, small outgrowths from the myotomes enter the limb rudiment to form the muscles of the fins. In tetrapods, however, the limb muscles develop from the same mass of mesenchyme that gives rise to the skeleton. Thus the muscles of the body and the muscles of the limbs have different origins—the first develop from the myotomes (thus from the somites), and the second develop from the lateral plate mesoderm via the limb buds.

Limb buds

The nerves supplying the limbs grow into the limb rudiments from the spinal cord and the spinal ganglia. The nerves are guided in some way by the limb rudiments, for, if limb rudiments are displaced by transplantation to

an abnormal position, the nerves still find their way and establish normal relationships to the limb muscles. Limb rudiments transplanted to sites very far from their normal positions induce local nerves to enter the limb, thereby making it motile.

EXCRETORY ORGANS

Nephrotome development

The kidneys of vertebrates consist of a mass of tubules that develop from the stalks of somites called nephrotomes. In some primitive vertebrates such as cyclostomes, the nephrotome in each segment gives rise to only one tubule (Figure 22), but, in the great majority of vertebrates, mesenchyme from adjacent nephrotomes fuses into a common mass that differentiates into a number of nephric tubules irrespective of the original segmentation of the mesoderm. Under primitive conditions each tubule opens by a funnel (the nephrostome) into the coelomic cavity; the opposite ends of the tubules fuse to form the collecting ducts of the kidney. A collection of capillaries (the glomerulus) becomes associated with the nephric tubule, forming its filtration apparatus. The glomerulus may be situated in the coelomic cavity opposite the nephrostome or, in all the more advanced animals, intercalated into the nephric tubule, forming with the latter a renal corpuscle of the kidney. In adults of all vertebrates above the amphibians, the nephrostomes disappear (or are never formed), so that the tubule begins with the renal corpuscle. Parts of the kidney in vertebrates can be distinguished as the pronephros (most anteriorly, at the forelimb level), the mesonephros (in the midtrunk region), and the metanephros (in the pelvic region). The three sections of the kidney develop at different stages, starting with the pronephros and ending with the metanephros. In their morphology and mode of development, the anterior parts show more primitive conditions than the posterior ones. The pronephros, developing early in embryo formation, is the functional kidney of fish and amphibian larvae. Its collecting duct opens into the hindmost part of the intestine, called the cloaca, and later also serves as the collecting duct of the mesonephros. In reptiles, birds, and mammals, the pronephros is nonfunctional, although even in these animals its duct persists as the mesonephric duct. The mesonephros develops later and replaces the pronephros as the functional kidney of adult fishes and amphibians and of the embryos of reptiles, birds, and mammals. The tubules of the mesonephros link up with the duct derived from the pronephros. The pronephric duct in fact stimulates the development of mesonephric tubules, and, in its absence, the mesonephros does not develop at all.

Development of the metanephros

The metanephros is found only in reptiles, birds, and mammals. It replaces the mesonephros of the early embryonic stages and continues as the functional kidney in the postembryonic and adult life of these animals. The metanephros develops from mesenchyme derived from the nephrotomes of the posterior part of the trunk and lying dorsal to the mesonephric duct. The actual differentiation is initiated by a dorsal outgrowth of the mesonephric duct, called the ureteric bud. The ureteric bud grows in the direction of the mesenchyme and becomes the ureter. Having penetrated the mass of mesenchyme, it starts to branch, producing the collecting tubules of the kidney; the mesenchyme, meanwhile, in response to the influence of the duct and its branches, aggregates to form the excretory tubules of the kidney. The influence of the ureter is indispensable for the development of the metanephric excretory tubules, for, if the ureter fails to develop or, in its outgrowth, stops short of reaching the kidney-producing mesenchyme, no kidney develops.

CIRCULATORY ORGANS

The rudiment of the heart in vertebrates develops from the ventral edges of the mesodermal mantle in the anterior part of the body, immediately adjoining the pharyngeal region. A group of mesodermal cells breaks away from the ventral edge of the lateral plate, takes a position just underneath the pharyngeal endoderm, and becomes arranged in the form of a thin-walled tube, which will become the endocardium, or lining of the heart (Figure 23). In vertebrates with complete cleavage, the endocardial tube is single and

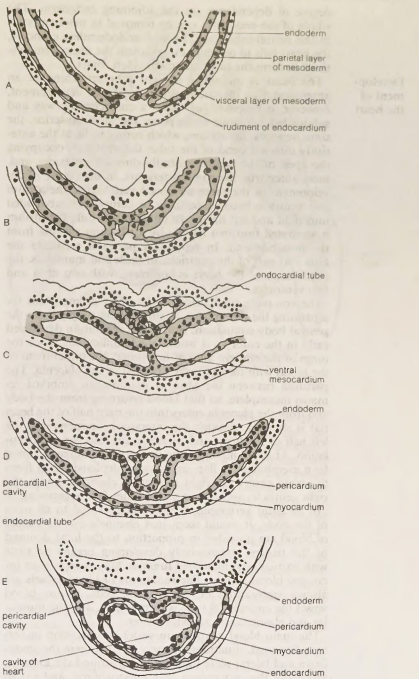


Figure 23. Development of the heart in amphibian embryos.
From O. Hertwig, *Handbuch der Vergleichenden und Experimentellen Entwicklungslehre der Wirbeltiere* (1906). Gustav Fischer Verlag.

medial from its start. In higher vertebrates with meroblastic cleavage—reptiles, birds, and mammals—the embryo in early stages of development is flattened out on the surface of the yolk sac; therefore, what are morphologically the ventral edges of the mesodermal mantle lie far apart on the perimeter of the blastodisc. As a result of this arrangement, two endocardial tubes are formed, one on either side of the embryo. Subsequently, when the embryo becomes separated from the yolk sac, the two endocardial tubes meet in the midline ventral to the pharynx and fuse, producing a single heart rudiment. After the formation of the endocardium, or the lining of the heart, the coelomic cavity in the lateral plate mesoderm adjoining the heart rudiment expands slightly and envelops the endocardial tube or tubes (Figure 23). The heart muscle layer, or myocardium, develops from the visceral (splanchnic) layer of the lateral plate that is in contact with the endocardial tube; the parietal (somatic) layer of the lateral plate forms the pericardium, or covering of the heart. The portion of the coelom surrounding the heart becomes separated from the rest of the body cavity and develops into the pericardial cavity.

The endocardial tube branches anteriorly into two tubes, the ventral aortas; a similar branching of the endocardial tube posteriorly forms the two vitelline veins, which carry blood from the midgut endoderm or from the yolk sac (when present) to the heart.

In its earliest development, the heart rudiment shows a

degree of dependence on the adjoining endoderm. The whole of the endoderm can be removed in newt embryos in the neural-tube stage. In such endodermless embryos, the heart fails to develop, even though the mesoderm destined to form the heart rudiment is left intact.

Development of the heart

The heart is initially a straight tube stretching in an anteroposterior direction. Rather early in development, however, it becomes twisted in a characteristic way and subdivided into four main parts: the most posterior, the sinus venosus; the atrium, which comes to lie at the anteriorly directed bend of the tube; the ventricle, occupying the apex of the posteroventrally directed inflexion; and, most anteriorly, the conus arteriosus. In the course of development in the more advanced vertebrates, the atrium and ventricle become partially or completely subdivided into right and left halves. In amphibians, only the atrium is separated into two halves, by a partition starting from the posterior end. In reptiles, a partition separates the atria and part of the ventricle. In birds and mammals, the subdivision of the heart is complete, with two atria and two ventricles.

The complete subdivision of the heart is important for separating the pulmonary, or lung, blood supply from the general body circulation. But, if this separation developed early in the embryo, it would create difficulties, since the lungs of the embryo are not functional; the enrichment of the blood with oxygen occurs instead in the placenta. The partition between the atria in mammalian embryos remains incomplete, so that blood returning from the body and from the placenta enters into the right half of the heart but is shunted (through the interatrial foramen) into the left half of the heart and thence again into general circulation. At birth, however, the interatrial foramen is closed by a membraneous flap, and oxygen-depleted blood from the body enters the right atrium, is channelled into the right ventricle, and thence to the lungs for oxygenating.

In an adult vertebrate, blood vessels extend to all parts of the body. It would seem that channels for the supply of blood are provided in proportion to the local demand of the tissues; progressively developing organs or parts with particularly intensified function always receive an increased blood supply. The rudiments of blood vessels are always aggregations of mesenchyme cells. In any blood vessel the endothelial tube is formed first, and the muscular and elastic layers are added later.

The main blood channels in vertebrates develop in certain favoured situations; namely (1) between the endoderm and lateral plate mesoderm; (2) around the kidneys, especially the pronephros and mesonephros; and (3) in connection with the heart, which is a special case of the first category.

From the paired forward extensions from the heart, the ventral aortas, loops develop between the pharyngeal clefts.

The aortic arches

These are the aortic arches (Figure 24), which served originally to supply blood to the gills in aquatic vertebrates. The arches are laid down in all vertebrates, six or more being found in cyclostomes and fishes; six are present in the embryos of tetrapods, but the first two are degenerate. The arches of the third pair develop as the carotid arteries, supplying blood to the head. Those of the fourth pair (and, exceptionally, in urodeles also the fifth) join dorsally to form the dorsal aorta, providing blood to most of the body. These are the systemic arches. The arches of the sixth pair are the pulmonary arches; in embryos they carry blood to the dorsal aorta, as well as to the lungs, but in fully developed amniotes (reptiles, birds, and mammals), they carry blood only to the lungs.

The paired posterior extensions of the heart of the early embryo are the vitelline veins, whose branches spread out between the lateral plate mesoderm and the endoderm, especially the endoderm of the yolk sac, when present. On their way to the heart, the vitelline veins pass through the liver and break up into a system of small channels—the hepatic sinusoids. Parts of the vitelline veins lying posterior to the liver become the hepatic portal veins, which carry blood from the intestine to the liver; the parts of the vitelline veins anterior to the liver become the hepatic veins, which carry blood fom the liver to the sinus venosus in lower vertebrates (anamniotes), but become the anterior section of the postcaval vein in amniotes.

Whereas the vitelline veins and, later, the hepatic portal vein carry blood from the endodermal parts of the embryo and from the yolk sac to the heart, the blood from the mesodermal and ectodermal parts is returned to the heart through a system of cardinal veins (Figure 25). These latter veins start their development in the form of an irregular

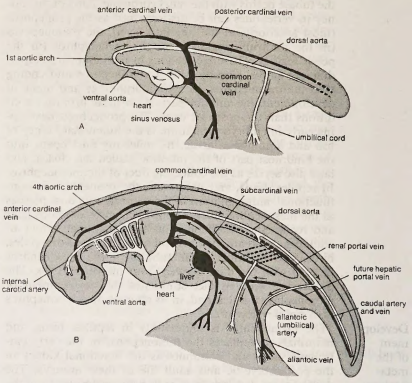


Figure 25: General arrangement of blood vessels in (A) an early amniote embryo and (B) the further development of the circulatory system. The arteries are shown in white, the veins in black.

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970). W.B. Saunders Company, Philadelphia

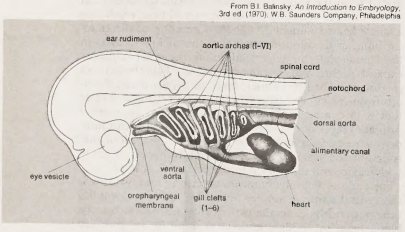


Figure 24: Relation of the aortic arches and the branchial clefts in a dogfish embryo.

sinus around the pronephros, connected by the common cardinal veins (ducts of Cuvier), on either side, to the sinus venosus. Extensions anteriorly and posteriorly give rise to the precardinal and postcardinal veins, respectively. The postcaval vein, present in terrestrial vertebrates, is a late acquisition, both in evolution and in embryogenesis; it is a result of the intercommunication of several venous channels, including the anterior portion of the vitelline veins.

Development of the cardinal veins

The first blood cells in vertebrate embryos form in association with the intestinal endoderm on the yolk sac. Groups of mesoderm cells derived from the splanchnic layer of the lateral plate (extra-embryonically in cases in which a yolk sac is present) become so-called blood islands, which are particularly conspicuous on the yolk sac of bird embryos (in the area vasculosa). In bird's eggs, the internal cells of the blood island start producing hemoglobin (gas-carrying component of blood) and become the first red blood cells (erythrocytes) as early as the second day of incubation. The outer cells of the blood islands develop into an endothelial layer and form a network of blood vessels covering part of the surface of the yolk sac. The network

acquires a connection to the vitelline veins and vitelline arteries (the latter being branches of the dorsal aorta); thus the blood corpuscles formed in the blood islands can enter the general blood circulation.

At later stages of embryogenesis, blood-cell formation shifts from the blood islands to the liver and, still later, to the bone marrow.

The lymphatic system, in a manner similar to the blood vessels, develops by the local aggregation of connective tissue to form lymphatic vessels.

REPRODUCTIVE ORGANS

In considering the development of reproductive organs, distinctions must be made between: (1) the origin of sex cells (gametes), (2) the origin and differentiation of the sex glands, or gonads (ovaries and testes), and (3) the origin and development of the supporting parts of the reproductive system (*e.g.*, genital ducts, copulatory organs).

Origin of germ cells

The germ (germinal) cells, which eventually give rise to the gametes, are often segregated from the somatic, or body, cells at a very early stage—during cleavage and before the subdivision of the embryo into ectoderm, mesoderm, and endoderm. In the invertebrate nematodes, the very first of these primordial germ cells is identifiable after as few as five divisions of the egg cell. The germ cell retains the large chromosomes present in the fertilized egg; in the somatic cells the chromosomes become fragmented. Subsequently, the single germ cell gives rise, by mitotic divisions, to all the gametes in the gonad.

In vertebrates, primordial germ cells arise outside the gonads, but they cannot be distinguished in early cleavage stages. In amphibians, cytoplasm at the vegetal pole, rich in ribonucleic acids, becomes incorporated into a number of cells, which, during cleavage and gastrulation, lie among the yolky endoderm cells. Later they migrate into the mesodermal layer and become incorporated into the rudiments of the gonads. In higher vertebrates, primordial germ cells can be recognized in the extra-embryonic endoderm of the yolk sac. In mammals, these cells subsequently migrate into the mesoderm and are located in the gonad rudiments. The mouse embryo, for example, originally has fewer than 100 primary germ cells; during their migration, however, their numbers increase as a result of repeated divisions, to 5,000 or more in the gonads.

Although the primordial germ cells either may appear before the separation of germinal layers or be found originally in the endoderm, the gonads are invariably of mesodermal origin. In vertebrates, the first trace of gonad development is a thickening of the coelomic lining on either side of the dorsal mesentery and medial to the kidney rudiments. The thickening, elongated anteroposteriorly, is known as the germinal ridge (Figure 26). The ridge protrudes into the coelomic cavity, and the fold of thickened epithelium becomes filled with mesenchyme. At this stage the primordial germ cells invade the rudiments of the gonads and become associated with the somatic cells of the germinal ridge. In the functionally differentiated gonads, only the actual gametes and their predecessors (spermatogonia and oogonia) are derived from the primary germ cells; the supporting cells are somatic cells of local mesodermal origin. In the ovaries, the follicle cells surrounding and nourishing the young egg cells (oocytes) are of somatic origin, as are also the connective tissue and blood vessels of the gonad. In the testes, supporting elements called Sertoli cells are somatic cells, as are the interstitial cells, which are scattered between the sperm-carrying tubules of the testes and believed to be the source of male hormones.

In the early stages of their development—even while the gonad rudiment is being invaded by primordial germ cells—the female and male gonads are in an indifferent stage. Only later does tissue differentiation of the gonads begin and male or female gonadal development proceed.

Development of genital ducts

The genital ducts, by which the eggs and sperm are carried away from the gonads, are, in vertebrates, linked with the excretory system. In the male, the seminiferous tubules connect with the nephric tubules of the mesonephros, and the sperm are carried to the exterior by way of the mesonephric duct. In males of lower vertebrates, the mesonephric duct thus serves as a channel

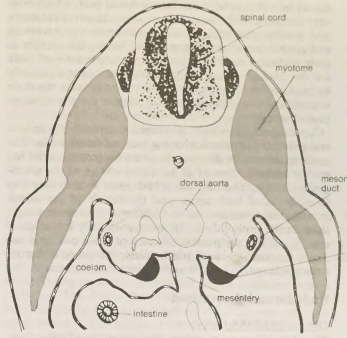


Figure 26: Transverse section of a mouse embryo showing the position of the germinal ridges and their relation to the mesonephric rudiment and the dorsal mesentery.
From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970), W.B. Saunders Company, Philadelphia

both for urine and for sex cells. In amniotes the development of the metanephros as the urine excreting organ has freed the mesonephric duct to carry products associated only with reproduction. In the female, a separate duct, the paramesonephric duct (Müllerian duct), develops beside the mesonephric duct. At its anterior end it utilizes the funnels of the pronephric tubules as its entrance (ostium). The paramesonephric duct develops initially in both female and male embryos. The ducts remain in an indifferent stage longer than the gonads. Eventually the sex hormones produced by the differentiating gonads cause a corresponding differentiation of the ducts. The mesonephric ducts, which become reduced in female embryos, remain in male embryos as ducts for conveying sperm (ductus deferens). The paramesonephric ducts, on the other hand, degenerate in male embryos but become the oviducts in female embryos (Figure 27). In mammals, the terminal portions of the paired oviducts differentiate as two uteri, which, in primates and man, fuse to form a single uterus.

In all terrestrial vertebrates except the placental mammals, the genital ducts, as well as the ducts of excretory organs, open into the cloaca. In mammals, however, the

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970), W.B. Saunders Company, Philadelphia

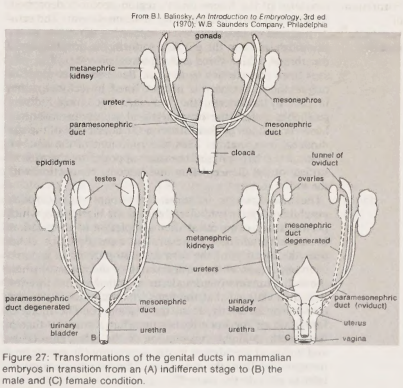


Figure 27. Transformations of the genital ducts in mammalian embryos in transition from an (A) indifferent stage to (B) the male and (C) female condition.

cloaca becomes subdivided into a dorsal part, which conveys the feces, and a ventral part, which receives excretory and genital products. In male mammals the excretory and genital ducts remain connected, having the urethra as their common outlet; in females the urethra serves only for the passage of urine and the uterus opens separately by means of the vagina. In nearly all vertebrates, the male nephric duct is utilized in some degree for the conduction of sperm.

Copulatory organs have developed independently in several groups of vertebrates having internal fertilization. The penis in mammals develops from an outgrowth called the genital tubercle, located at the anterior edge of the urinogenital orifice. The tubercle is laid down in a similar way in embryos of both sexes, and the region of the urinogenital orifice remains in an indifferent state even longer than do the genital ducts. In a comparatively late stage of embryonic life the genital tubercle of male embryos encloses the urethral canal and becomes the penis; in female embryos it remains small and becomes the clitoris.

Endodermal derivatives

THE ALIMENTARY CANAL

The alimentary canal is the chief organ developing from endoderm. The way it forms depends on the type of egg cleavage. In eggs with holoblastic (complete) cleavage, after gastrulation the invaginated mass of endoderm lines the archenteron, the cavity of which becomes the alimentary canal, or gut. In eggs with meroblastic (partial) cleavage—and also in mammals (despite their complete cleavage)—the endoderm is produced in the form of a sheet lying flat over the yolk-sac cavity. Subsequently, folds of endoderm and splanchnic mesoderm appear—first anteriorly, then laterally, and lastly posteriorly—and sink, converging ventrally under the embryo and cutting off the future gut cavity from the cavity of the yolk sac. The most anterior and posterior portions of the gut separate, but the middle part remains in open communication with the yolk sac throughout embryonic life, eventually becoming reduced to the yolk stalk, which passes through the umbilical cord.

The alimentary canal of vertebrates becomes differentiated into the oral cavity, pharynx, esophagus, stomach, and intestine. Whether derived from an archenteron or formed by folding of the endodermal sheet, the canal initially does not possess an opening at its anterior end. This is also the case in some lower chordates and echinoderms, which are grouped together with vertebrates as the Deuterostomia, or animals with secondary mouths.

Formation of vertebrate mouth

In vertebrates, a mouth forms by a rupture at the anterior end, where the endoderm is in contact with ectoderm. The ectoderm of the future mouth region becomes depressed, forming a mouth invagination, or stomodaeum. The ectodermal and endodermal layers separating the cavity of the stomodaeum from the gut fuse to form the oropharyngeal membrane, which thins and ruptures, providing free passage from the exterior to the gut. Because of its mode of origin, the oral cavity is in part lined by ectoderm and in part by endoderm, the two parts becoming indistinguishable. Before the oropharyngeal membrane ruptures, however, a small pocket forms on the dorsal side of the stomodaeal invagination. This, the rudiment of the anterior lobe of the hypophysis, becomes apposed to the ventral surface of the diencephalon and loses its connection with the mouth cavity.

The anal opening in some exceptional cases (urodele amphibians) is derived directly from the blastopore, which persists as a narrow canal after completion of gastrulation. In other vertebrates, however, the anus develops either near the location of the former blastopore or in a corresponding region at the posterior end of the embryo, where the last remnants of mesoderm migrated to the interior. It is thus claimed that the anus in vertebrates is derived, directly or indirectly, from the blastopore. The mode of formation of the opening is somewhat similar to that of the mouth. A slight invagination of the ectoderm occurs, and a cloacal membrane forms, separating the ectodermal invagination from the gut cavity. The membrane ruptures later to provide the anus.

THE PHARYNX AND ITS OUTGROWTHS

The anterior portion of the endodermal gut, lying immediately posterior to the mouth cavity, expands laterally as the pharynx. The lateral pockets of the pharyngeal cavity, called the pharyngeal pouches, perforate the mesodermal layer, reach the ectoderm, and break through to form pharyngeal, or gill, clefts (Figure 24). In fishes and larvae of amphibians, these clefts develop gills and become respiratory organs. Pharyngeal pouches develop in the early embryos of all vertebrates, including the airbreathing terrestrial reptiles, birds, and mammals. The number of pouches has been reduced in the course of evolution from six or more to four in tetrapods, and the posterior pouches may not actually break through.

The consistent development of pharyngeal pouches and clefts indicates their importance in vertebrate development. Many parts of the vertebrate body are derived from, or dependent on, the pharyngeal pouches; for example, the aortic arches—the most important blood vessels of a vertebrate—develop between successive pharyngeal pouches (Figure 25). Skeletal visceral arches also occur between consecutive pharyngeal pouches (they do not develop if the pharyngeal pouches are prevented from developing). In adult terrestrial vertebrates, parts of the visceral arches are transformed into the hyoid apparatus, supporting the tongue, the auditory ossicles, and parts of the larynx and trachea. Furthermore, some of the material of the pharyngeal pouches is utilized for the formation of the parathyroid glands and the thymus; the former are indispensable glands of internal secretion, and the latter are a source, in mammals, of cells that produce antibodies. The pharynx also produces the rudiment of the thyroid gland as a ventral outgrowth.

THE LIVER, PANCREAS, AND LUNGS

Three additional important organs develop from the endoderm: the liver, the pancreas, and the lungs. The liver develops as a ventral outgrowth of the endodermal gut just posterior to the section that eventually will become the stomach. Initially, the liver takes the form of a tubular gland, but it soon acquires a close relationship to the blood sinuses and capillaries, forming lobules around blood vessels rather than around glandular ducts. The pancreas develops from three independent rudiments: two ventral ones, formed just posterior to the liver rudiment, and a dorsal one. The ventral and the dorsal rudiments fuse in most vertebrates to form one organ with a complicated system of ducts opening into the duodenum, a portion of the small intestine. The lungs develop from a ventral hollow outgrowth of the gut, which is located just posterior to the pharyngeal region; the outgrowth branches into a right and left trunk that grow posteriorly beside the esophagus and then expand into hollow sacs, in lower terrestrial vertebrates, or into a system of tubes, in birds and mammals.

Development of lungs

The endodermal parts of the alimentary system are, along their entire length, encased by the splanchnic mesoderm of the lateral plates. The coelomic cavities of the right and left sides fuse ventral to the gut but remain separated dorsally by their respective walls, which form the dorsal mesentery—a double membrane by which the gut is suspended from the dorsal side of the body cavity and through which blood vessels and nerves reach the gut (Figure 22). The layer of splanchnic mesoderm next to the endoderm produces the connective tissue and muscular layers of the gut. During development of the glands of the alimentary canal (*e.g.*, pancreas, salivary glands), the mesoderm forms a connective tissue capsule around the branching tubules of the gland. The development of the tubules is dependent on this mesodermal capsule and cannot proceed without it.

Postembryonic development

After partially developing within the egg membranes or within the maternal body, the newly formed individual emerges. The new animal is then born (ejected from the mother's body) or hatched from the egg. The condition of the new organism at the time of birth or hatching

differs in various groups of animals, and even among animals within a particular group. In sea urchins, for example, the embryo emerges soon after fertilization, in the blastula stage. Covered with cilia, the sea-urchin blastula swims in the water and proceeds with gastrulation. Frog embryos emerge from the egg membranes when the main organs have already begun to develop, but functional differentiation of the tissues is unfinished; for instance, the components of the eyes and ears are far from complete, the mouth is not yet open, and the gut is filled with yolk-laden cells. Certain birds (called precocial) emerge from the egg covered with downy feathers and can run about soon after hatching, whereas others (altricial) hatch naked, with only rudiments of feathers, and are quite unable to move around. Among mammals there is a great range in the degree of development at birth. In marsupials, such as opossums and kangaroos, the young are born incompletely developed and very small; the young are then kept for a long time in the pouch of the mother, all the while firmly attached to the teats and suckling. Many small mammals are helpless at birth. Mice are born naked and blind; puppies and kittens are born covered with fur but with unopened eyes. Newborn human babies have their eyes open but cannot move themselves about for several months. Hoofed mammals, on the other hand, bear young that can stand up and run after their mothers within a few hours of birth.

Egg tooth of birds

In birds the hard shell is broken by the hatchling's beak, which is provided with a sharp tubercle on its top. A similar "egg tooth" appears on the tip of the snout of hatchling reptiles. Many arthropods have a preformed line of fragility that allows part of the eggshell to be burst open like a lid, allowing the young to emerge. Birth in mammals is effected through the contraction of smooth muscles of the uterus.

THE LARVAL PHASE AND METAMORPHOSIS

The organism emerging from the egg or from the maternal body, apart from being incompletely developed, may have an organization more or less different from that of an adult. In some cases the difference is so great that, without knowing the origin of the eggs or without following the young through their full course of development, it would be impossible to know that the young and the adult are of the same animal species. Such young, called larvae, transform into the adult form by a process of metamorphosis. The term larva also applies to young that resemble the adult form but differ from it in some substantial respect, as in possessing organs not present in the adult or in lacking an important structure (apart from sex glands and associated parts, which tend to develop later in life in most animals). Larvae in different animals have special names given to them, such as the tadpole of frogs, the caterpillar of butterflies, and the fry of fishes.

Advantages of larvae

The development of the embryo into a larva rather than directly into an organism similar to the adult has various advantages. At the time of emergence from the egg, the new individual is relatively small, and the organization that enables the adult to lead a particular mode of life may not be suitable for a miniature copy of the adult. The larva may have to procure food for itself and, being small, may not be able to feed in the same way as the adult. It also may not be able to use effectively the same defense mechanisms the adult possesses. The larval stage enables an animal to avoid such hazards; it provides a mode of life and corresponding organization better suited to the smaller size of the newly emerged organism. Another advantage is that the larva may be able to exploit an entirely different environment because its organization is very different from that of the adults. A terrestrial adult may have aquatic larvae, a flying adult may have burrowing larvae, and a parasitic adult may have a free-living larva. A third advantage of a larval stage emerges in animals whose adult stages are sessile or restricted in their movements; the larvae can move freely, either of their own accord or on water currents. In this way the larvae of sedentary animals serve for the dispersal of the species. Lastly, the larval stage is of great advantage for certain internal parasites, which, once inside a host, cannot transfer to another. New hosts are infected instead by the larval stages. (The usual means of attaining this end is for the parasite to produce enormous quantities of eggs and rely on the passive entry of the eggs into the new host with food. A more efficient way, however, is for a mobile larva to enter the new host actively.)

A large number of marine invertebrates possess floating larvae that have hairlike projections (cilia) as their means of locomotion. There are three main types of larvae, characteristic of large subdivisions of the animal kingdom.

The planula larva of coelenterates has an elongated shape and cilia covering its entire surface. The internal organization is simple, hardly beyond differentiation into ectoderm and endoderm in the interior. The larva does not feed but serves only for dispersal.

The trochophore larva

The trochophore larva is found in many marine invertebrates. Typically, as in polychaetes, it has an alimentary canal with mouth and anus and a ring of ciliated cells arranged anterior to the mouth. It also possesses a sensory organ and rudiments of mesoderm. Cilia around the mouth bring in food—unicellular plants and other small particles. The larva thus not only serves for dispersal but also feeds and grows before it transforms into an adult worm. Other trochophore larvae are found in marine mollusks and in certain marine worms. The larva of echinoderms is similar to the trochophore in possessing a gut and a ciliary band, but the arrangement of the latter is different. The echinoderm larva also feeds and grows as well as serves for dispersal.

Larvae of very different kinds are found in many arthropods. In crustaceans the larva, called nauplius, does not differ substantially in mode of life or means of locomotion from the adult but has fewer appendages than the adult. A typical crustacean nauplius has three pairs of legs and an unpaired simple eye. Additional pairs of appendages and paired compound eyes appear in the course of a sometimes prolonged development. In insects the larva differs from the adult by the absence of wings but, in addition, may have a different mode of life and different way of feeding. Among chordates the tunicates (sea squirts) deserve attention; the larval form is a free-swimming creature, showing unmistakable relation to vertebrates, but the adult is sedentary, with much reduced nervous and muscular systems. The tadpole of a frog differs from the adult in being totally aquatic, in possessing a tail and gills for respiration, and in having a mouth adapted for feeding on plants. The adult frog is adapted to land life, except for reproductive periods, has no tail and no gills, and is an active predator.

Types of metamorphosis

Metamorphosis, the transformation of the larva into an adult, is a more or less complicated process depending on the degree of difference between the two forms. The transformation may be gradual, extend over a long period, and involve a number of intermediate stages; alternatively, the transformation may be achieved in one step. In the latter case, especially if the difference between the larva and adult is great, large parts of the body of the larva, including all the specifically larval organs, disintegrate (necrobiotic metamorphosis). At the same time, organs of the adult are built up, sometimes from reserve groups of cells that remain undifferentiated or nonfunctional in the larva. A good illustration of the distinction between gradual and abrupt metamorphosis occurs among the insects. In more primitive insects, such as cockroaches and grasshoppers, metamorphosis is gradual. The larva, often referred to as a nymph, has more or less the same organization as the adult, or imago; it feeds in a similar way but differs from the adults in lacking wings and in having incomplete sex organs. The wings appear in later stages of larval life; they are small at first but increase with each molt, and they attain full size and functional capacity at the last (imaginal) one. The larva of other insects, such as beetles, butterflies, and wasps, is a grub or caterpillar, a wormlike creature not even remotely resembling the adult. The difference in organization is so profound that the transformation cannot be achieved gradually, and an intermediate resting, or pupal, stage is interposed between the larva and imago. The pupa neither feeds nor moves, as the larval organs inside are destroyed and replaced with organs of the adult.

The nymph stage

including wings and sex organs. Eventually, when formation of the adult organs is complete, the pupal skin is cast off, and the adult emerges. The destruction of the larval parts may be far reaching and include even the skin and most of the alimentary canal. The tissues of the adult are formed from groups of reserve cells that were present all along in the larva as imaginal disks.

Necrobiotic metamorphosis is observed in the tunicate larva, in which the tail, including notochord, nerve cord, and muscles, and most of the brain, including eye and statocyst, are destroyed at the same time that the large pharyngeal cavity of the adult develops. A tadpole metamorphosing into an adult frog loses its tail—the cells of which are destroyed and devoured by phagocytic cells—its gills, and its larval mouthparts; concurrently the legs of the adult frog develop progressively, the structure of the mouth and alimentary canal change, and the skin acquires a bony (keratinized) layer and a system of subcutaneous glands (Figure 28).

From E. Witschi, *Development of Vertebrates* (1956). W.B. Saunders Company, Philadelphia

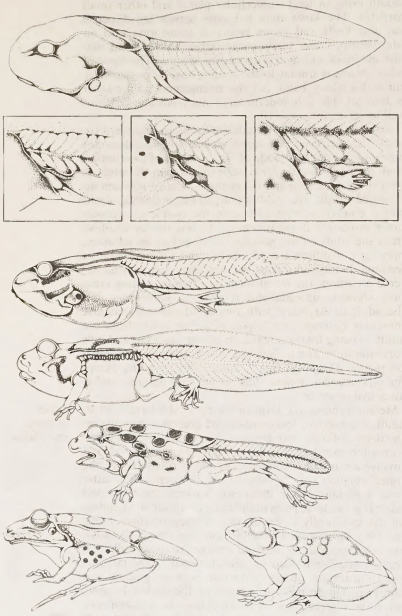


Figure 28: Stages in the metamorphosis of the frog from aquatic tadpole to terrestrial adult frog, including details of development of the hindleg.

The complicated changes taking place during metamorphosis, especially in the case of necrobiotic metamorphosis, must be performed in a coordinated way. So that no changes are made prematurely and no organ systems are left behind in the general transformation, some common signal for the change must be provided. For both insect and amphibian metamorphoses, which have been the most extensively studied, the signal is a hormonal one, sent in the blood to all the cells and tissues of the body.

Molting and hormonal function

Metamorphosis in an insect is complicated by the fact that the rigid cuticle covering its body is very restrictive; new features can appear only after a molt, when the old cuticle is replaced by a newly formed one. Molting in insects is caused by the action of two hormones. In the brain of insects, several groups of neurosecretory cells produce the first hormone. This brain hormone does not itself affect molting but stimulates the prothoracic gland, a loose mass of secretory cells situated in the thorax in close association with tracheal tubes. In response to the stimulation by the brain hormone, the prothoracic gland releases into the blood a second hormone, the molting hormone, or ecdysone. Under the influence of ecdysone, the tissues of the body produce a new cuticle under the old one, after which the old cuticle is shed (the actual molting). The new cuticle embodies any new developmental features that were scheduled to appear. The kind of feature that emerges after a molt is controlled by a third organ of internal secretion, the corpus allatum, secretory tissue situated posterior to the brain, near or around the dorsal aorta and usually appearing as a pair of separate or fused organs. The corpora allata emit the juvenile hormone, which, as long as it circulates in the blood, acts to perpetuate the larval form. As the larva approaches the end of its development, however, the corpora allata stop producing juvenile hormone or reduce its quantity; whereupon, the larva, at the next molt, metamorphoses into an adult. Withdrawal of the juvenile hormone is the immediate cause of metamorphosis, in conjunction with the brain hormone and ecdysone, which are responsible for the shedding of the larval cuticle and for the production of the new cuticle embodying the features of the imago. Metamorphosis through the stage of the pupa is effected by diminishing levels of juvenile hormone, which determine first the transformation of the larva into a pupa and, with further reduction of the juvenile-hormone level, the final step of transformation of the pupa into the adult.

The metamorphosis of a tadpole into a frog also depends upon two hormones: one initiating the process and the other directly influencing the tissues involved in the change. The first hormone is the thyrotropic hormone, produced by the hypophysis. It has no immediate effect on the tissues of the body but activates the thyroid gland to produce several substances, the most important of which is thyroxine. Thyroxine and other iodine-containing compounds circulate in the blood and cause changes that, in their entirety, constitute the process of metamorphosis. It is remarkable that different tissues react in different ways to the presence of thyroxine. The muscles of the tadpole's tail degenerate, whereas the muscles of the trunk and legs are not affected; in fact, the growth and development of limbs are stimulated as a part of metamorphosis. The effect of the hormone depends on the nature of the reacting cells and tissues—*i.e.*, on their competence—just as the embryonic inductor in the earlier stages of development influences only cells with the competence for a particular kind of reaction.

DIRECT DEVELOPMENT

If an animal after birth or emergence from an egg differs from the adult in comparatively minor details (apart from not having functional sex organs), the development is said to be direct. There is no larval stage and no metamorphosis. Direct development does not mean, however, that no changes occur between birth and adulthood. One very obvious change is the growth of the animal.

The rate of growth—not absolute increase—is highest in the early stages of postembryonic life; subsequently, growth continues to slow, ceasing completely at the attainment of adulthood. The rate of growth is dependent on many factors, both external (feeding, temperature) and internal. Of the internal factors, the most important are hormones, especially the growth hormone produced by the hypophysis. If the growth hormone is produced in insufficient quantities, the result is dwarfism; if it is produced in excessive quantities, the result is gigantism.

Factors influencing rate of growth

In the case of direct development, the most important change is the attainment of sexual maturity, which is achieved in several steps and involves the action of several hormones. The gonad rudiments and rudiments of the supporting parts of the reproductive system remain

inactive long after birth. At the approach of adulthood, however, two sets of hormones come into action: hypophyseal hormones stimulate the gonads to activity, and gonadal hormones (produced by the gonads) cause the supporting sex organs and other sex characters to become fully developed. To become functional, the gonads must be acted upon by secretions from the hypophysis. In immature females the follicle-stimulating hormone, which alone causes the egg follicles and the oocytes to grow, and the luteinizing hormone stimulate the follicle cells to produce the female sex hormone, estrogen, which effects the development of the uterus, the milk glands, and other characteristics of the female sex. In the male, the same hypophyseal hormones are produced, with the result that the testes start to produce sperm and to secrete the male sex hormone, androgen. It appears that the luteinizing hormone is the more active in the male sex, being able to cause both spermatogenesis and androgen secretion. Androgen, in turn, brings about the development of the penis, the descent of the testicles before birth, the appearance of typical male hair growth, and other secondary sex characteristics.

Maturity and death

Sexual maturity and the ensuing reproductive activity mark the pinnacle of development and morphogenesis and, for many animals, herald the end of life. The biological goal of the entire process is achieved with the launching of the next generation, and the life cycle that runs from the formation of gametes by one generation to the formation of gametes by the next generation is completed. In many animals the females die after laying their eggs; the males may have died earlier, after pairing. Indeed, some males (spiders, praying mantises) are eaten by the females immediately after copulation.

The developmental period can only truly be said to end with the termination of an organism, for much activity continues to unfold new developmental sequences, not all of them progressive and favourable, to be sure. Senescence, or a decline in abilities, signals advancing age in mammals but is not a general occurrence in the animal kingdom. Far more animals continue to function at near-peak capacity well into old age. And even among those species—salmons, eels, many moths—whose members die after a single reproductive act, death is relatively swift and not accompanied by a prolonged period of deterioration.

In most animals the reproductive potential is not exhausted in a single act of gamete production, but the sexually mature individuals remain alive and reproduce repeatedly. In these cases life may extend long beyond the first attainment of reproductive ability and be accompanied by further growth of the individuals, as occurs in most fishes, amphibians, and reptiles, and also in mollusks and certain other invertebrates. In the case of prolonged life-spans, however, reproductive activity may cease with advancing age, and a senile involution take place, as is observed mainly in mammals and, particularly, in man. The changes taking place may be described as regressive development. In most animals, however, the end of life is not preceded by any overt traces of senility. As a general rule, then, the attainment of reproductive ability may be said to be the final phase of progressive development among animals.

A gradual loss of alertness and vigour is typical of the aging pattern of primates and is especially important to man. (For additional information, see below *Aging and senescence*.) (B.I.B.)

HUMAN GROWTH AND DEVELOPMENT

The human body, like that of most animals, develops from a single cell produced by the union of a male sex cell and a female sex cell. Human development follows closely the basic vertebrate pattern, and it departs only in certain details from the type specifically characteristic of mammals (see above *Animal development*). A prenatal period, in which most of the developmental advances occur, is followed by a long postnatal period. Only at about the age of 25 are the last progressive changes completed.

Human embryology: early stages

FROM FERTILIZATION TO PLACENTATION

Penetration of the ovum by the sperm and their union

Fertilization. The development and liberation of the male and female gametes are steps preparatory to their union through the process of fertilization. Random movements first bring some spermatozoa into contact with follicle cells adhering to the secondary oocyte, which still lies high in the uterine tube. The sperm then propel themselves past the follicle cells and attach to the surface of the gelatinous zona pellucida enclosing the oocyte. Some sperm heads successfully penetrate this capsule by means of an enzyme, hyaluronidase, which they secrete, but only one makes contact with the cell membrane and cytoplasm of the oocyte and proceeds farther. This is because the invading sperm head releases a substance that initiates surface changes in the oocyte cytoplasm that other competitors cannot master.

The sole successful sperm is engulfed by a conical protrusion of the oocyte cytoplasm and is drawn inward. Once within the periphery of the oocyte, the sperm advances toward the centre of the cytoplasm; the head swells and converts into a typical nucleus, now called the male pronucleus, and the tail detaches. It is during the progress of these events that the oocyte initiates its final maturation division. Following the separation of the second polar body, the oocyte nucleus reconstitutes typically and is then called the female pronucleus of the ripe egg. It is now ready to unite with its male counterpart and thereby consummate the total events of fertilization.

The two pronuclei next approach, meet midway in the egg cytoplasm, and lose their nuclear membranes. Each resolves its diffuse chromatin material into a complete, single set of 23 chromosomes. Centrioles (complex particles involved in cell division), apparently supplied by the sperm, appear, and a mitotic spindle organizes with the two sets of chromosomes arranged midway on it—ready to proceed with a typical mitosis. This climax in the events of fertilization creates a joint product named the zygote. It contains all the essential factors for the development of a new individual.

The fundamental results of fertilization are the following: (1) reassociation of a male and female set of chromosomes (thus restoring the full number and providing the basis for biparental inheritance and for variation); (2) establishment of the mechanism of sex determination for the new individual (this depending on whether the male set of chromosomes included the X- or the Y- chromosome); (3) activation of the zygote, initiating a beginning toward the production of a new individual.

Cleavage and blastulation. *Cleavage.* The onset of mitosis (ordinary cell proliferation by division) in the activated zygote is a first step toward development in the ordinary sense of that term, and the cells so produced are the first external sign of future body building. To this end, the relatively enormous zygote directly subdivides into many smaller cells of conventional size, suitable as early building units for the future organism. This process is called cleavage and the resulting cells are blastomeres (Figures 29A–D). The tendency for the progressive increase in cell numbers to follow a doubling sequence is soon disturbed and then lost. Each blastomere receives the full complement of paternal and maternal chromosomes.

Formation of blastomeres

Subdivision of the zygote into blastomeres begins while it is still high in the uterine tube. The cohering blastomeres are transported downward chiefly, at least, by muscular contractions of the tubal wall. Such transport is relatively rapid until the lower end of the tube is reached, and here cleavage continues for about two days before the multicellular cluster is expelled into the uterus. The full reason for

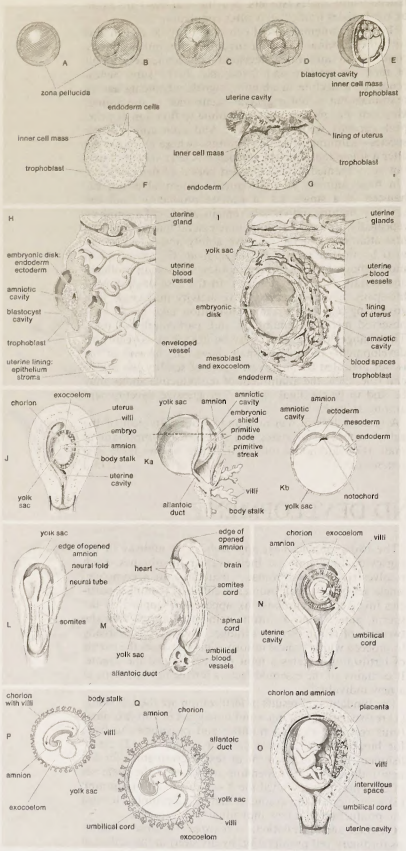


Figure 29: *Development of the human embryo.* A–E (rabbit) and F–G (monkey) illustrate comparable first stages of human development. (A–D) Cleavage of ovum (magnified 80×). (E) Early free blastocyst, partially opened (magnified 80×). (F) Older free blastocyst, halved (magnified 60×). (G) Late attached blastocyst, halved (magnified 70×). (H–J) Half sections of embryos implanted in uterine lining: (H) at 7½ days (magnified 80×); (I) at 13 days (magnified 37×); (J) at 23 days (magnified ¼×). (Ka) Embryo of 18 days at disk (or shield) stage, three-quarter view (magnified 9×). (Kb) Cross section through same embryo at level of broken line in Ka. (L) Embryo of 21 days with amnion opened, back view (magnified 8×). (M) Embryo of 23 days with yolk sac and opened amnion (magnified 9×). (N,O) Embryo of six weeks and fetus of three months, respectively, within halved amnion, chorion, and uterus (magnified ¼×). (P,Q) Diagrams showing especially amnion growth and formation of umbilical cord; stages correspond to J and N.
Drawing by James F. Didusch

this delay is not clear, but it serves to retain the cleaving blastomeres until the uterine lining is suitably prepared to receive its prospective guest.

Since the human egg contains little inert yolk material, and since this is distributed rather evenly throughout the cytoplasm, the daughter cells of each mitosis are practically equal in size and composition. This type of cleavage is known as total, equal cleavage. The sticky blastomeres adhere and the cluster is still retained for a time within the gelatinous capsule—the zona pellucida—that had enclosed the growing and ovulated oocyte. There is no growth in the rapidly dividing blastomeres, so that the total mass of living substance does not increase during the cleavage period. Cleavage is not a mechanism designed primarily to distribute particular developmental qualities, either nuclear or cytoplasmic, to specific blastomeres, which then carry out assignments that give rise to particular parts of the embryo. In a sense, therefore, cleavage is essentially a nonspecific fractionating preliminary rather than a process of truly constructive development.

Morula and blastocyst. By the fourth day after ovulation, a cluster of about 12 blastomeres passes from the uterine tube into the uterus. When the cluster numbers 12 to 16 blastomeres it is called a morula (Figure 29D). By the time some 30 blastomeres have been produced, pools of clear fluid accumulate between some of the internal cells, and these spaces soon coalesce into a common subcentral cavity. The resulting hollow cellular ball is a blastula of a particular type that occurs in mammals and is called a blastocyst; its cavity is the blastocoel (Figure 29E, F).

Passage of blastomeres into uterus; formation of morula and blastocyst

An internal cellular cluster, eccentric in position and now named the inner cell mass, will develop into the embryo. The external capsule of smaller cells, enveloping the segregated internal cluster, constitutes the trophoblast. It will contribute importantly to the formation of a placenta. During its stay within the uterine cavity, the blastocyst loses its gelatinous capsule, imbibes fluid, and expands to a diameter of 0.2 millimetre (0.008 inch); this is nearly twice the diameter of the zygote at the start of cleavage. Probably several hundred blastomeres have formed before the blastocyst attaches to the uterine lining.

Implantation and placentation. Late in the sixth day after ovulation the naked, sticky blastocyst comes into contact with the uterine lining and adheres to it (Figure 29G). The site of attachment is variable and not predetermined. The uterine lining has already been preparing, under the influence of ovarian hormones, for the reception of the blastocyst. Among these preparations has been the elaboration and expulsion, by the uterine glands, of a secretion that serves as nourishment for the blastocyst, both when it is free and during its implantation. Directly after blastocyst attachment come its establishment within the thickened uterine lining and the participation of its trophoblastic capsule in the differentiation of a placenta, a structure that enables the developing embryo to enter into a prompt, physiological dependence on the mother.

Implantation. The trophoblast of the blastocyst exerts an enzymic, destructive influence on the swollen uterine lining, leading to erosion of both the superficial epithelium of the uterine lining and also its deeper, cellular connective tissue (Figure 29H). This early stage of invasion ends in a few days; the blastocyst is then completely buried within a more superficial and compact layer of the total uterine lining (Figure 29H, I). While the blastocyst is completing this phase of implantation, its original shell of cellular trophoblast steadily proliferates a multitude of cells that lose their outermost membranes and merge. The result is a thick peripheral layer consisting of a common mass of cytoplasm in which many nuclei are embedded. This external investment is called syncytial trophoblast.

The implanted blastocyst next proceeds to establish itself as a parasite within the uterus. The syncytial trophoblast becomes a spongy shell containing irregular cavities (Figure 29I). This expanding mass destroys cellular connective tissue and glands encountered in its path. Both the cellular and derivative syncytial trophoblast have the capacity of ingesting such tissue. The erosive process also taps capillaries and small blood vessels, and blood liberated in this way is likewise taken up into the trophoblast. Such en-

Syncytial trophoblast

gulfed blood cells and fragments of tissue are presumably utilized for nourishment in the early period of establishment of the expanding blastocystic sac. Yet these erosive activities decline in intensity by the end of the third week of development, and at this time the sac is completing the first phase of its specialization (Figure 29L).

Ectopic pregnancy

Occasionally a fertilized egg fails to reach the uterus, implanting and beginning to develop elsewhere. This outcome is called an ectopic or extra-uterine pregnancy. The commonest ectopic site is the uterine tube, but the peritoneum lining the abdominal cavity and even the interior of the ovary are also involved, though rarely. The unsuitability of all these sites for continued development usually leads to early death and resorption of the embryo.

Placentation. The irregular strands of invasive syncytial trophoblast constitute a first stage in the formation of true villi (which form part of the placenta and are briefly described below). Primitive connective tissue soon lines the interior of the blastocyst wall, and this complex (trophoblast and connective tissue) is then named the chorion. Connective tissue promptly grows into the trophoblastic strands, and blood vessels develop in the tissue. The result is the production of many chorionic villi, each resembling a tiny, branching bush (Figure 29J, Ka).

In the fourth week of development, the essential arrangements have been established that make possible those physiological exchanges between mother and fetus that characterize the remainder of pregnancy. The deepest embedded portion of the chorionic wall becomes the so-called chorionic plate of the developing placenta. From the plate extend the main stems of chorionic villi, which give off progressively subdividing branches (Figure 29N, O). In general, the side branches are free, whereas apical ones tend to attach to the maternal tissue and serve as anchors. The villous trees occupy a labyrinthine space between the villi that was created by erosion of the uterine lining. Trophoblast not only covers the chorionic plate and its villi but also spreads like a carpet over the eroded surface of the maternal tissue.

Tapped uterine arteries open into the trophoblast-lined intervillous space, their blood bathing the branches and twigs of the villous trees. This blood drains from the intervillous space through similarly tapped veins. Arterial blood of the embryo, and later fetus, passes through vessels of the umbilical cord to the chorionic plate. Thence it is distributed to the villous stems, branches, and twigs through vessels in their connective-tissue cores. Return of this blood to the fetus is by a reverse route.

The circulation of maternal blood through the intervillous space is wholly separate from fetal blood coursing through the chorion and its villi. Communication between the two is solely by diffusive interchange. The barrier between the two circulations consists of the trophoblastic covering of villi, the connective tissue of the villous cores, and the thin lining of the capillaries that are contained in the villous cores. The placenta serves the fetus in several ways, most of which involve interchanges of materials carried in the bloodstreams of the mother and fetus. These functions are of the following kinds: (1) nutrition, (2) respiration, (3) excretion, (4) barrier action (*e.g.*, prevention of intrusions by bacteria), and (5) synthesis of hormones and enzymes. (For functional details and for a description of the mature, expelled placenta, see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Human reproduction from conception to birth: The normal events of pregnancy.*)

Functions of the placenta

Extra-embryonic membranes. The way in which the encapsulating membrane of the blastocyst becomes the chorion, and the most deeply embedded part of it becomes the fetal placenta, has already been described. There are still other important membranes that develop from those portions of the inner cell mass of the blastocyst that are not directly involved in becoming an embryo.

Yolk sac. Cells split off from the inner cell mass of the blastocyst and fashion themselves into a primitive yolk sac (Figure 29J, Ka). The roof of the sac then folds into a tubular gut, whereas the remainder becomes a vascularized bag that attains the size of a small pea (Figure 29M). In other vertebrates, such as amphibians and birds, the yolk sac is large and contains a store of nutritive yolk. But in man and other true mammals there is practically none. A slender neck, the yolk stalk, soon connects the rapidly elongating gut with the fast growing yolk sac proper. The stalk detaches from the intestine early in the second month, but the shrunken sac commonly persists and can be found in the expelled afterbirth.

Amnion. A cleft separates the outermost cells of the inner cell mass of the blastocyst from the remainder, which then becomes the embryonic disk (Figure 29H). The split-off, thin upper layer is the amnion, which remains attached to the periphery of the embryonic disk. As the disk folds into a cylindrical embryo, the amniotic margin follows the underfolding, and its line of union becomes limited to the ventral (frontward) body wall, where the umbilical cord attaches (Figure 29P, Q). The amnion becomes a tough, transparent, nonvascular membrane that gradually fills the chorionic sac and then fuses with it (Figure 29N, O). At the end of the third month of pregnancy, the nonplacental extent of this nearly exposed double membrane comes into contact with the lining of the uterus elsewhere. Fusion then obliterates the uterine cavity, which has been undergoing progressive reduction in size. For the remainder of pregnancy the only cavity within the uterus is that of the fluid-filled amniotic sac.

The amniotic sac

Clear, watery fluid fills the amniotic sac. The embryo is suspended in this fluid and thus can maintain its shape and mold its body form without hindrance. Throughout pregnancy the amniotic sac serves as a water cushion, absorbing jolts, equalizing pressures, and permitting the fetus to change posture. At childbirth it acts as a fluid wedge that helps dilate the neck of the uterus. When the sac ruptures, about a quart of fluid escapes as the "waters." If the sac does not rupture or if it covers the head at birth, it is known as a caul.

Allantois. The allantois, a tube of endoderm (the innermost germ layer), grows out of the early yolk sac in a region that soon becomes the hindgut. The tube extends into a bridge of mesoderm (the middle germ layer) that connects embryo with chorion and will become incorporated into the umbilical cord (Figure 29P, Q). The human allantoic tube is tiny and never becomes a large sac with important functions, as it does in reptiles, birds, and many other mammals. In the second month it ceases to grow, and it soon is obliterated. Blood vessels, however, develop early in its mesodermal sheath, and these spread into the chorion and vascularize it. Throughout pregnancy they will keep the embryo in close relationship with the mother's uterine circulation.

Umbilical cord. As the ventral body wall closes in, the yolk stalk and allantois are brought together, along with their mesodermal sheaths and blood vessels (Figure 29P, Q). Enclosing everything is a wrapping of amnion. In this manner a cylindrical structure, the umbilical cord, comes to connect the embryo with the placenta (Figure 29N, O). It will serve the embryo and fetus as a physiological lifeline throughout the period of pregnancy. The mature cord is about 1.3 centimetres (0.5 inch) in diameter, and it attains an average length of nearly 50 centimetres (two feet).

FORMATION OF THE THREE PRIMARY GERM LAYERS

The inner cell mass, attached to the deep pole of the implanted blastocyst, is sometimes called the embryoblast, since it supplies the materials used in the formation of an embryo. The cellular mass flattens and enters into the process of gastrulation, through which the three primary germ layers segregate and the gastrula stage, the next advance after the blastula, begins to take form. First, cells facing the cavity of the blastocyst arrange into a layer named the endoderm (Figure 29H, I). The thick residual layer, temporarily designated as epiblast, is the source of a definitive uppermost sheet, the ectoderm, and an intermediate layer, the mesoderm. In this second phase of gastrulation, some cells of the epiblast migrate to the midline position, then turn downward and emerge beneath as mesoderm. Such cells continue to spread laterally, right and left, between the endoderm and the residue of epiblast, which is now definitive ectoderm (Figure 29Kb).

Gastrulation

The site where the migratory mesodermal cells leave the epiblast is an elongated, crowded seam known as the

primitive streak (Figure 29Ka). Similar migrating cells produce a thick knob at one end of the primitive streak. Their continued forward movement from this so-called primitive knot produces a dense band that becomes the rodlike notochord.

The germ layers are not segregated sheets whose cells have predetermined, limited capacities and inflexibly fixed fates in carrying out organ-building activities. Rather, the layers represent advantageously located assembly grounds out of which the component parts of the embryo emerge normally, according to a master constructional plan that assigns different parts to definite spatial positions and local sites. Thus, although the germ layers have developmental potencies in excess of their normal developmental fates, their ordinary participation in organ forming does not deviate from a definite, standard program.

The derivatives of the primary germ layers can best be presented in tabular form. In naming the germ-layer origin of an organ, only the principal functional tissue is designated (Table 1). In a few instances, such as the suprarenal (adrenal) glands and the teeth, a compound organ has important parts of different origin.

Table 1: Derivatives of Primary Germ Layers

ectoderm	mesoderm	endoderm
epidermis	epithelium of:	epithelium of:
cutaneous derivatives	circulatory system	pharynx
epithelium of:	spleen; lymph nodes	thyroid; thymus
mouth; oral glands	urogenital system	parathyroid
nasal passages	body cavities	digestive tube;
sense passages	connective tissues;	liver; pancreas
central nervous system	blood; bone marrow	larynx; trachea; lungs
peripheral nervous system	muscular tissues	urinary bladder; urethra
hypophysis;	skeletal tissues	vestibule; vagina
suprarenal medulla	suprarenal cortex	

GROWTH AND DIFFERENTIATION

Growth. Growth is an increase in size, or bulk. Cell multiplication is fundamental to an increase in bulk but does not, by itself, result in growth. It merely produces more units to participate in subsequent growing. Growth is accomplished in several ways. Most important is synthesis, by which new living matter, protoplasm, is created from available foodstuffs. Another method utilizes water uptake; a human embryo of the early weeks is nearly 98 percent water, while an adult is 70 percent fluid. A third method of growth is by intercellular deposition; cells manufacture and extrude such nonliving substances as jelly, fibres, and the ground substance of cartilage and bone. Through these activities a newborn baby is several thousand million times heavier than the zygote from which it came.

Ways in which growth is accomplished

It is obvious that uniform growth throughout the substance of a developing organism would merely produce a steadily enlarging spherical cellular mass. Local diversities in form and proportions result from differential rates of growth that operate in different regions and at different times. The particular program of starting times and growth rates, both externally and internally in the human embryo, constitutes its characteristic growth pattern. Abnormal growth occurs occasionally, and growth may be excessive or deficient. Also, such departures may be general or local, symmetrical or asymmetrical. General gigantism usually starts before birth, and the oversized baby continues to grow at an accelerated rate. On the other hand, an existing hereditary predisposition may not be aroused into action until some time during childhood. In a reverse manner, general dwarfism may exist before birth, the individual continuing to grow slowly and stopping at the usual time. Another type is normal in size at birth, grows normally for a while, and then comes to a premature arrest.

Differentiation. In a developing organism, differentiation implies increasing structural and chemical complexity. One kind of differentiation concerns changes in gross shape and organization. Such activities, related to molding the body and its integral parts into form and pattern, comprise the processes called morphogenesis. The processes of morphogenesis are relatively simple mechanical acts: (1) cell migration; (2) cell aggregation, forming masses, cords, and sheets; (3) localized growth or retardation, resulting, in enlargements or constrictions; (4) fusion; (5) splitting, including separation of single sheets into separate layers, formation of cavities in cell masses, and forking of cords; (6) folding, including circumscribed folds that produce inpocketings and outpocketings; (7) bending, which, like folding, results from unequal growth.

Morphogenesis and histogenesis

A second kind of differentiation refers to progressive changes occurring in the substance and structure of cells, whereby different kinds of tissues are created. These changes, and the synthetic processes underlying them, constitute histogenesis. The zygote contains all the essential factors for development, but they exist solely as an encoded set of instructions localized in the genes of chromosomes and bearing no direct physical relationship to the future characteristics of the developing embryo. During histogenesis these instructional blueprints are decoded and transformed, through cytoplasmic syntheses, into the several types and subtypes of tissues that are the structural and functional units of organs. At first, the cells of each germ layer lack an identifiable shape and are similar in chemical composition, but an invisible type of chemical differentiation soon enters. After the elaboration of specific enzyme patterns and syntheses, certain groups of cells progressively assume distinctive characters that permit their fates to be recognized. Such early stages in definite lines of differentiation of cells are often designated by the suffix "-blast" (*e.g.*, myoblast; neuroblast).

The emerging cell types are discrete entities, without intermediates; for example, a transitional form between a muscle cell and a nerve cell is never seen. Neither can different, local parts of a cell carry out different types of tissue specialization, such as nerve at one end and muscle at the other end; nor can a cell, once fully committed to a particular type of specialization, abandon it and adopt a new course.

"Dedifferentiation"; the four types of tissue

Under certain conditions, differentiated cells may reverse their course and return to a simpler state. Thus, under a changed environment, cartilage may lose its matrix, and its cells may come to resemble the more primitive tissue from which it arose. Nevertheless, despite such reversal and apparent simplification ("dedifferentiation"), these cells retain their former histological specificity. Under suitable environmental conditions they can differentiate again but can only regain their previous definitive characteristics as cartilage cells.

The final result of histogenesis is the production of groups of cells similar in structure and function. Each specialized group constitutes a fundamental tissue. There are four main types of such tissues: each of the three germ layers gives rise to sheetlike epithelia, which cover surfaces, line cavities, and are frequently glandular; ectoderm also forms the nervous tissues; mesoderm also produces the muscular tissues; it also differentiates into blood and the fibrous connective tissues (including two further specialized types, cartilage and bone).

EMBRYONIC ACQUISITION OF EXTERNAL FORM

At the end of the second week, the embryonic region is a nearly circular plate within its well-embedded, differentiating chorionic sac (Figure 29I). This embryonic disk consists of two layers (epiblast and endoderm), with a third layer (mesoderm) just starting to spread between them. A hollow, dome-shaped amnion sac attaches to the margin of the upper layer of the disk, and a hollow yolk sac is similarly continuous with the lower layer. A broad cellular bridge attaches the complex to the chorion.

Early in the third week the embryonic disk has enlarged and become pear-shaped in outline, and a well-formed primitive streak occupies the midline of its caudal (hind) half, which is narrower. Cells from the epiblast are passing through the streak and spreading laterally in both directions beneath the uppermost layer, now ectoderm (Figure 29Ka, Kb). In this way the embryonic disk becomes three layered, and the gastrula stage of development comes to an end. At the middle of the week a thickening, the head process, is extending forward from a knoblike primitive knot located at the head end of the primitive streak. These linear thickenings define the median plane of the future

embryo and thus divide the embryonic disk into precise right and left halves.

Toward the end of the week the disk elongates and becomes slipper-shaped in outline; a slight constriction demarcates it from the attached yolk sac. Growth has lengthened the region ahead of the now receding primitive streak. Here, in the midline, the ectoderm bears a definite gutter-like formation called the neural groove; it is the first indication of the future central nervous system. Beneath the groove the mesodermal head process presently rounds into an axial rod, the notochord, that serves as a temporary "backbone." By the end of the third week a head fold, paired lateral body folds, and a tail fold become prominent, demarcating a somewhat cylindrical embryo from the still broadly attached yolk sac (Figure 29L). The neural folds, flanking the neural groove, converge and begin to meet midway of their lengths, thereby producing a neural tube at that level. Mesoderm, alongside the notochord, begins to subdivide into paired blocks called somites, and the outlines of the somites show externally. From them, muscles and vertebrae will differentiate later. This stage, when the embryo is fashioning a neural tube, is often designated as a neurula.

In the fourth week the embryo goes beyond the external characteristics of vertebrates in general and becomes recognizable as a mammal. The week is marked by profound changes, during which the embryo acquires its general body plan. There is an increase in total length from about two to five millimetres, but size is quite variable among smaller specimens. Better correlated with the degree of development is the number of mesodermal somites, which attain their full number of about 42 during the fourth week. Some of the head of an early embryo arises from the embryonic disk in front of the primitive knot. But as the primitive streak shortens and its caudal retreat continues, such structures as the neural tube and notochord are added progressively in the wake of that retreat, and additional somite pairs also appear in steady succession.

The most important manoeuvre in the establishment of general body form is the transformation of the flat embryonic disk into a roughly cylindrical early embryo, which is attached to the yolk sac by a slender yolk stalk (Figure 29M). Three factors cooperate in producing this change: (1) There is more rapid expansion of both the embryonic area and the yolk sac than in the region joining the two. The enlarging embryonic area at first buckles upward

Factors in formation of early embryo

By courtesy of the Carnegie Institution of Washington

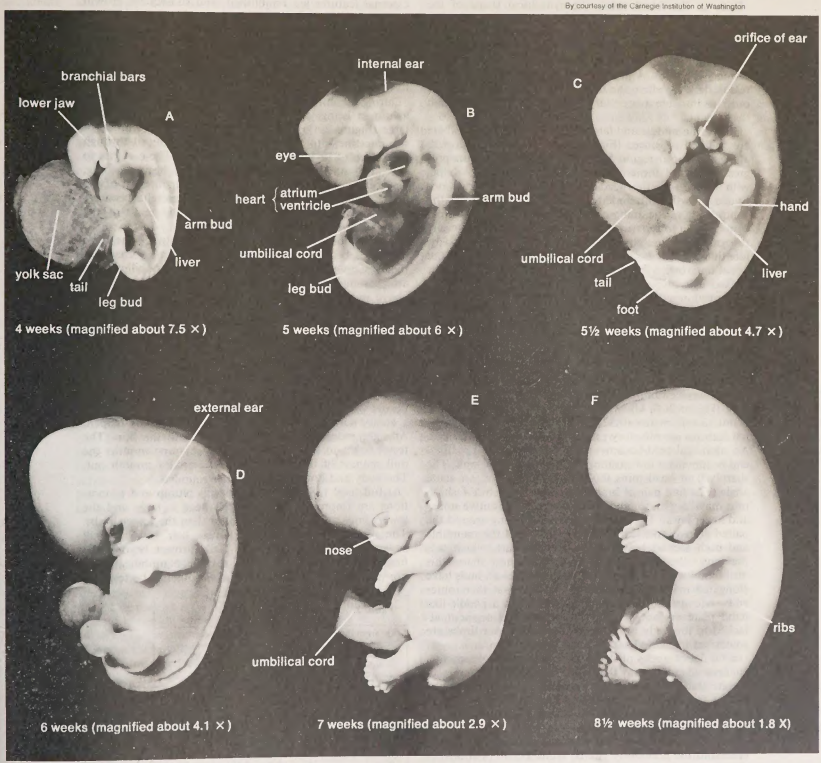


Figure 30: Human embryos from four to 8½ weeks.

and then overlaps the more slowly growing margin. Since growth is particularly rapid at the future head end and tail end, the embryo becomes elongate. (2) In conjunction with this overgrowth there is important underfolding, again most pronounced at the front and hind ends. Underfolding is produced by differential elongation in the regions of the brain and tail bud. Conspicuous is the change in the future cardiac (heart) and foregut region, which swings beneath the brain as on a hinge. (3) A certain amount of true constriction, through growth, purses all of these parts at the site of the future umbilicus.

Throughout the entire period when the body and its parts are being laid down, developmental advances tend to appear first at the head end and then progress tailward. For this reason, many structures that extend along the body for a distance show a gradation in development. The size advantage gained initially by the head end of the embryo is relinquished very slowly. Even in an infant the relatively large head and long arms are striking. A further tendency toward progressively graded development occurs from the middorsal line in a lateral (sideward) and then ventral (frontward) direction. All such relations are the visible expressions of stages in growth and differentiation.

Early in the fourth week the cylindrical shape of the embryo is plain, even though the folding-off process is far from complete (Figure 29M). The neural tube is still open near both ends, and at the head end the broader neural folds indicate the future brain and even its three primary divisions. A pronounced bulge beneath the brain region denotes where the heart is forming precociously in order to institute a necessary, prompt circulation through the placenta.

During the middle and late days of the fourth week there are marked advances (Figure 30A). Accelerated growth along the dorsal region bends the total body length progressively until the embryo assumes a striking C-shape, with the tips of the head and tail not far apart. Continued growth and underfolding close in much of the ventral side of the embryo, so that a free head and upper trunk, and a lower trunk and a prominent tail, are easily recognizable. Forebrain, midbrain, and hindbrain can be identified, largely because of a sharp bend in the midbrain. Local outgrowths from each side of the forebrain produce stalked eye cups, and a pair of inpocketings of the ectoderm alongside the hindbrain sink beneath the surface as otic vesicles, forerunners of the inner ears. Bulges indicative of the heart and liver are prominent. Formations called branchial arches, reminiscent of the gill arches of fishes and aquatic amphibians, become conspicuous in the future jaw–neck region. Paired swellings ("buds") off the trunk foretell the locations of the upper and lower limbs.

Changes in the fifth to eighth weeks

A five-week embryo is about eight millimetres (0.32 inch) long (Figure 30B, C), whereas at six weeks the length is about 13 millimetres (0.52 inch) (Figure 30D). New external features are olfactory pits at the tips of the bent head. An umbilical cord becomes a definite entity, its proximal end occupying a low position on the abdominal wall. The sharply bent head joins the rest of the body at an acute angle. The first pair of branchial arches branch Y-fashion into maxillary and mandibular processes (primitive upper and lower jaws). The external ears are forming around the paired grooves located between each half of the mandible and each second branchial arch. The heart, which was previously the chief ventral prominence, now shares this distinction with the rapidly growing liver. Limb buds have elongated markedly and become flattened at their outer ends. A constriction on each bud separates a paddle-like hand plate or foot plate from a cylindrical segment attached to the body wall. Predictably, the upper limbs are somewhat further advanced than the lower pair.

In the latter weeks of the second month, developmental changes advance from those that distinguish primates to a state that is recognizably human (Figure 30E). At the end of the second month (30 millimetres', or 1.2 inches', length) the stage of the embryo ends; henceforth, until birth, "fetus" is the preferred term.

The fetal stage

In the seventh and eighth weeks the head becomes more erect and the previously curved trunk becomes straighter (Figure 30E, F). The heart and liver, which earlier dominated the shape of the ventral body, yield to a more evenly rounded chest–abdomen region. The tail, which at an earlier time was one fifth of the embryo's length, becomes inconspicuous both through actual regression and through concealment by the growing buttocks. The face rapidly acquires a fairly human appearance; eyes, ears, and jaws are prominent. The eyes, previously located on the sides of the head, become directed forward; the nose lacks a bridge and so is of the "pug" type, with the nostrils directed forward instead of downward. A mandibular branch of each Y-shaped branchial arch combines with its mate to form the lower jaw. The maxillary branch on each side joins an elevation on the medial (inner) side of the corresponding nostril to produce the more complicated upper jaw. Branchial arches, other than those forming the jaws and external ears, are effaced through incorporation into an emerging, recognizable neck. Limbs become jointed, and the earlier hand plates and foot plates differentiate terminal digits. Primitive external genitalia appear, but in a nondistinctive, sexless condition.

Almost all of the internal organs are well laid down at the end of eight weeks, when the embryo is little more than 25 millimetres (one inch) long. The characteristic external features are established, and subsequent growth merely modifies existing proportions without adding new structure. Similarly, the chief changes undergone by internal organs and parts are those of growth and tissue specialization. At eight weeks the neuromuscular mechanism attains a degree of perfection that permits some response to delicate stimulation.

Development from the third month to birth

During the third month the young fetus clearly resembles a human being, although the head is disproportionately large (Figure 29O). The previous protrusion of much of the intestine into the umbilical cord is reduced through the return of its loops into the abdomen. The ears rise to eye level and the eyelids fuse shut. Nails begin forming; ossification (bone-forming) centres appear in most of the future bones; and the sex of external genitalia becomes recognizable. (In this paragraph, and in the next two, the months are lunar months, of 28 days).

At four months individual differences between the faces of fetuses become distinguishable. The face is broad but the eyes are now less widely separated. The umbilical cord attaches higher on the abdominal wall; this location is above an expanding region between the cord and the pubis (front bones of the pelvis) that scarcely existed previously.

At five months downy hairs (lanugo) cover the body, and some head hairs appear. The skin is less transparent. Fetal movements ("quickening") are felt by the mother. At six months eyebrows and eyelashes are clearly present. The body is lean, but its proportions have improved. The skin is wrinkled. At seven months the fetus resembles a dried-up old person. Its reddish, wrinkled skin is smeared with a greasy substance (vernix caseosa). The eyelids reopen. At eight months fat is depositing beneath the skin. The testes begin to invade the scrotum. At nine months the dull redness of the skin fades and wrinkles smooth out. The body and limbs become better rounded.

At full term (38 weeks) the body is plump and proportions are improved, although the head is large and the lower limbs are still slightly shorter than the upper limbs. The skin has lost its coat of lanugo hair, but it is still smeared with vernix caseosa. Nails project beyond the finger tips and to the tips of toes. The umbilical cord now attaches to the centre of the abdomen. The testes of

Table 2: Age, Size, and Weight

age from date of conception (calendar months)	sitting height		weight	
	mm	in.	g	lb
Two	28	1	2.25	.75 oz
Three	75	3	25	1 oz
Four	135	5.3	170	6 oz
Five	185	7.3	440	14 oz
Six	225	9	820	1.75 lb
Seven	270	10.6	1,380	3 lb
Eight	310	12.2	2,220	5 lb
Nine	360	14.2	3,150	7 lb

males are usually in the scrotum: the greater lips of the female external genitalia, which previously gaped, are now in contact. Cranial bones meet except at some angular junctions, or "soft spots."

The average time of delivery is 280 days from the beginning of the last menstrual period, whereas the duration of pregnancy (age of the baby) is about 266 days (38 weeks; Table 2). Pregnancy may extend to 300 days, or even more, in which case the baby tends to be heavier. Premature babies born under 27 weeks of age rarely survive, whereas those more than 30 weeks old usually do survive. (For additional information on the events of delivery, see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Human reproduction from conception to birth: Parturition.*)

Abnormal development

MULTIPLE BIRTHS

It is both unusual and abnormal for the human species to produce more than one offspring at a time. "Twins" and "twinning" are used as general terms for any number of multiple births, as the same basic principles apply.

Fraternal and identical twins

Fraternal twins stem from multiple ovulations in the same cycle. Each oocyte develops singly in a separate follicle, is shed and fertilized individually, develops within its own chorionic sac, and forms an individual placenta. In some instances, two blastocysts implant close together and the expanding placentas meet and fuse. In such double placentas, however, the two blood circulations rarely communicate. The word dizygotic technically designates two-egg twins. Such pairs obviously are independent in sex determination and bear no more resemblance than do other children of the same parents. Properly speaking, they are merely littermates. Nearly three-fourths of all American twins are dizygotic, whereas the Japanese ratio is only one-fourth. A tendency toward such multiple births exists in some family lines.

Wholly different are those true twins who are always of the same sex and are strikingly similar in physical, functional, and mental traits. Such close identity is enforced by their derivation from a single ovulated and fertilized egg, and hence by their acquisition of identical chromosomal constitutions. This twin type is named monozygotic. Three-fourths of such pairs develop within a common chorionic sac and share a placenta; one-fourth have individual sacs and placentas. The latter condition results from a mishap before implantation, when the cleavage cells separate into two groups and then become individually implanting blastocysts. There is no discernible hereditary tendency toward the production of monozygotic twins.

Several atypical processes involving the inner cell mass or embryonic plate can produce separate embryos within a single sac: (1) The inner cells of a blastocyst may segregate into two masses. (2) Somewhat later in time, two embryonic axes may become established on a single embryonic disk. (3) A single axis may subdivide by fission or budding. (4) Duplication of any sort may combine with secondary subdivision; the Dionne quintuplets are believed to have followed this sequence, which is also normal for the regular quadruplets of the Texas armadillo.

Conjoined twins

The abnormal nature of monozygotic twinning is emphasized when the two individuals are conjoined as a "double monster." The degree of union varies from slight to extensive, and the possession of a single or double set of internal organs depends on the intimacy of fusion at any particular level. Union occurs by the heads, upper trunks, or lower trunks; the joining may be by the dorsal, lateral, or ventral surfaces. Sometimes there is a marked disparity in the size of the two components, and in such instances the smaller twin is called a parasite. Conjoining results from divergent growth at the front or hind end of the emerging primitive axis of an embryo, or at both ends.

FETAL DEVIATIONS

Human embryos are subject to disease, abnormal development, and abnormal growth. Decline and death can occur at any stage, but since most deaths occur in the first two or three weeks of development they usually escape notice. Probably little more than half of all zygotes reach full-term birth. Most abnormalities resulting from faulty development originate within the first seven weeks of pregnancy, before the prospective mother is surely aware of her condition. Abnormalities that do occur in living infants tend toward the milder types, since the severe mishaps commonly terminate development before birth.

Folklore maintains that a pregnant woman may "mark" her baby through incurring physical injuries or becoming subjected to horrors or repulsive sights. As there are no nervous connections between mother and fetus, such beliefs lack foundation. Moreover, practically all of the alleged causative experiences occur long after the "related" abnormalities have been established in the embryo.

Defective health of the mother can, in some instances, become a cause of the physical impairment or death of a fetus. Certain infectious diseases, for example, may result in fetal injury; such related causative organisms can be a virus (German measles), a spirochetal microorganism (syphilis), or a protozoon parasite (toxoplasmosis). Also, placental disorders, malformations of the mother's reproductive organs, and inadequate functioning of her endocrines may provide an unfavourable environment for normal development. Birth itself imposes the risk of oxygen deficiency or other injury; either may result in some malfunctioning of the brain.

TERATOLOGY

Teratology is concerned with all features of abnormal generation and development of the embryo (embryogenesis) and their end products. The incidence of defective development is high. One infant in 14 that survive the neonatal period bears an abnormality of some kind and degree, and half of these babies have more than one malformation. Internal, concealed defects are more numerous than external ones, and some defects do not become apparent until childhood. One baby in 40 is born with a structural defect that needs treatment. Some types of abnormality are commoner in males (*e.g.*, pyloric stenosis, the narrowing of the opening between the stomach and the intestine), while other types predominate in females (*e.g.*, dislocated hip). Besides the obvious derangements of a gross nature, there are aberrations at the molecular level known as inborn errors of metabolism. In these an enzyme deficiency blocks the course of intermediary metabolism and results in abnormal chemical functioning. Such errors involve proteins, carbohydrates, lipids, and pigments. The abnormal products may be stored or excreted.

Incidence, causation, and types of abnormality

Important among causes of abnormalities are hereditary factors. Such include gene mutations, which may become Mendelian dominants (*e.g.*, fused fingers, which need be inherited from only one parent to appear in the offspring), recessives (*e.g.*, albinism, which does not become evident unless its gene is inherited from both parents), or sex-linked factors (*e.g.*, hemophilia). Besides the heritable defects, whose possibilities of recurrence can be estimated, there are many genetic results that are due to chance, are not passed on, and do not occur in other offspring. An unequal distribution of chromosomes during meiosis, leading to abnormal assortments, occurs in somatic (non-sex) chromosomes (*e.g.*, Down's syndrome) and in sex chromosomes (*e.g.*, Klinefelter's syndrome).

Environmental factors, both external and internal, are also important. Among physical agents, mechanical pressures or blows are no longer considered to be significant because of the protection supplied by the uterus and the fluid-filled amniotic sac. On the other hand, irradiation is a wholly effective physical agency, as experiments have amply proved. Various chemical agents, used experimentally on pregnant animals, are highly teratogenetic (productive of physical defects within the uterus), and at least one drug (thalidomide) has demonstrated its potency on human embryos. Deficiencies of some fetal hormones are associated causally with gross bodily defects (*e.g.*, male hormone and false hermaphroditism, a condition in which the gonads are of one sex but some appearances suggest the other). Similarly, hormonal excess can be productive (*e.g.*, growth-promoting hormone and gigantism). Vitamin deficiencies are effective experimentally but are not factors in ordinary substandard maternal diets. The role of infectious agents has been men-

tioned previously. Immunity incompatibilities may exist between mother and fetus (*e.g.*, Rh blood factor and fetal hemolysis, the release of hemoglobin from red blood cells).

There appear to be several ways in which teratogenetic agents can affect susceptible embryonic cells. But whatever the manner of interference may be, the final result is probably either cell impairment or death or a changed rate of growth. Either of these measures puts local development out of step with adjoining parts and upsets the coordinated schedule of development.

Development of organs

ECTODERMAL DERIVATIVES

Origin of the skin

Integumentary system. The skin has a double origin. Its superficial layer, or epidermis, develops from ectoderm. The initial single-layered sheet of epithelial cells becomes multilayered by proliferation, and cells nearer the surface differentiate a horny substance. Pigment granules appear in the basal layer, at least, of all races. The epidermis of the palm and sole becomes thicker and more specialized than elsewhere. Cast-off superficial cells and downy hairs mingle with a greasy glandular secretion and smear the skin in the late fetal months; the pasty mass is called vernix caseosa. The deep layer of the skin, or dermis, is a fibrous anchoring bed differentiated from mesoderm. In the later fetal months the plane of union between epidermis and dermis becomes wavy. The permanently ridged patterns are notable at the surface of the palm and sole.

Nails, hairs, and skin glands

Nails develop in pocket-like folds of the skin near the tips of digits. During the fifth month specialized horny material differentiates in proliferating ectodermal cells. The resulting nail plate is pushed forward as new plate substance is added in the fold. Fingernails reach the finger tips one month before birth. Hairs, produced only by mammals, begin forming in the third month as cylindrical pegs that grow downward from the epidermis into the dermis. Cells at the base of the peg proliferate and produce a horny, pigmented thread that moves progressively upward in the axis of the original cylinder. This first crop of hairs is a downy coat named lanugo. It is prominent by the fifth month but is mostly cast off before birth. Unlike nails, hairs are shed and replaced periodically throughout life.

Sebaceous glands develop into tiny bags, each growing out from the epithelial sheath that surrounds a hair. Their cells proliferate, disintegrate, and release an oily secretion. Sweat glands at first resemble hair pegs, but the deep end of each soon coils. In the seventh month an axial cavity appears and later is continued through the epidermis. The mammary glands, peculiar to mammals, are specialized sweat glands. In the sixth week a thickened band of ectoderm extends between the bases of the upper and lower limb buds. In the pectoral (chest) region only, gland buds grow rootlike into the primitive connective tissue beneath. During the fifth month 15 to 20 solid cords foretell the future ducts of each gland. These later canalize and open into a pit that after birth elevates into a nipple. Until late childhood the mammary glands are identical in both sexes.

Mouth and anus. *Mouth.* The mouth is a derivative of the stomodaeum, an external pit bounded by the overjutting primitive nasal region and the early upper and lower jaw projections. Its floor is a thin membrane, where ectoderm and endoderm fuse. Midway in the fourth week this membrane ruptures, making continuous the primitive ectodermal mouth and endodermal pharynx (throat). Lips and cheeks arise when ectodermal bands grow into the mesoderm and then split into two sheets. Teeth have a compound origin: the cap of enamel develops from ectoderm, whereas the main mass of the tooth, the dentin, and the encrusting cementum about the root differentiate from mesoderm. The salivary glands arise as ectodermal buds that branch, bushlike, into the deeper mesoderm. Berrylike endings become the secretory acini (small sacs), while the rest of the canalized system serves as ducts. The palate is described in relation to the nasal passages. A tiny pocket detaches from the ectodermal roof of the stomodaeum and becomes the anterior, or frontward, lobe of the hypophysis, also called the pituitary. The anterior lobe fuses with the neural lobe of the gland (see below).

Anus. A double-layered, oval membrane separates the endodermal hindgut from an ectodermal pit, called the proctodaeum, the site of the future anal canal and its orifice, the anus. Rupture at eight weeks creates a communication between the definitive anus and the rectum.

Origin, histogenesis, and morphogenesis of central nervous system

Central nervous system. Both the brain and the spinal cord arise from an elongate thickening of the ectoderm that occupies the midline region of the embryonic disk. The sides of this neural plate elevate as neural folds, which then bound a gutter-like neural groove (Figure 29L). Further growth causes the folds to meet and fuse, thereby creating a neural tube. The many-layered wall of this tube differentiates into three concentric zones, first indicated in embryos of five weeks. The innermost zone, bordering the central canal, becomes a layer composed of long cells called ependymal cells, which are supportive in function. The middle zone becomes the gray substance, a layer characterized by nerve cells. The outermost zone becomes the white substance, a layer packed with nerve fibres. The neural tube also is demarcated internally by a pair of longitudinal grooves into dorsal and ventral halves. The dorsal half is a region associated with sensory functioning and the ventral half with motor functioning.

Neural tube layers

The gray substance contains primitive stem cells, many of which differentiate into neuroblasts. Each neuroblast becomes a neuron, or a mature nerve cell, with numerous short branching processes, the dendrons, and with a single long process, the axon. The white substance lacks neuroblasts but contains closely packed axons, many with fatty sheaths that produce the whitish appearance. The primitive stem cells of the neural tube also give rise to nonnervous cells called neuroglia cells.

Brain. The head end of the neural plate becomes expansive even as it closes into a tube. This brain region continues to surpass in size the spinal cord region. Three enlargements are prominent, the forebrain, midbrain, and hindbrain. The forebrain gives rise to two secondary expansions, the telencephalon and the diencephalon. The midbrain, which remains single, is called the mesencephalon. The hindbrain produces two secondary expansions called the metencephalon and the myelencephalon.

The telencephalon outpouches, right and left, into paired cerebral hemispheres, which overgrow and conceal much of the remainder of the brain before birth. Late in fetal life the surface of the cerebrum becomes covered with folds separated by deep grooves. The superficial gray cortex is acquired by the migration of immature nerve cells, or neuroblasts, from their primary intermediate position in the neural wall. The diencephalon is preponderantly gray substance, but its roof buds off the pineal body, which is not nervous tissue, and its floor sprouts the stalk and neural (posterior) lobe of the pituitary. The mesencephalon largely retains its early tubular shape. The metencephalon develops dorsally into the imposing cerebellum, with hemispheres that secondarily gain convolutions clothed with a gray cortex. The myelencephalon is transitional into the simpler spinal cord. Roof regions of the telencephalon, diencephalon, and myelencephalon differentiate vascular choroid plexuses (including portions of the pia matter, or innermost brain covering, that project into the ventricles, or cavities, of the brain), which secrete cerebrospinal fluid.

Spinal cord. For a time the spinal cord portion of the neural tube tapers gradually to an ending at the tip of the spine. In the fourth month it thickens at levels where nerve plexuses, or networks, supply the upper and lower limbs; these are called the cervical and lumbosacral enlargements. At this time the spine begins to elongate faster than the spinal cord. As a result, the caudal (hind) end of the anchored cord becomes progressively stretched into a slender, nonnervous strand known as the terminal filament. Midway in the seventh month the functional spinal cord ends at a level corresponding to the midpoint of the kidneys. Both the brain and the spinal cord are covered with a fibrous covering, the dura mater, and a vascular membrane, the pia-arachnoid. These coverings differentiate from local, neighbouring mesoderm.

Peripheral nervous system. In general, each craniospinal nerve has a dorsal (posterior) root that bears a ganglion (mass of nerve tissue) containing sensory nerve cells and

Histogenesis of peripheral nervous system

their fibres, and a ventral (anterior) root that contains motor nerve fibres but no nerve cells. Ganglion cells differentiate from cells of the neural crest, which is at first a cellular band pinched off from the region where each neural fold continues into ordinary ectoderm. Each of these paired bands breaks up into a series of lumps, spaced in agreement with the segmentally arranged mesodermal somites. Neuroblasts within these primordial ganglia develop a single stem and hence are called unipolar. From this common stem one nerve process, or projection, grows back into the adjacent sensory half of the neural tube; another projection grows in the opposite direction, helping to complete the dorsal root of a nerve. Neuroblasts of motor neurons arise in the ventral half of the gray substance of the neural tube. They sprout numerous short, freely branching projections, the dendrons, and one long, little-branching projection, the axon; such a neuron is called multipolar. These motor fibres grow out of the neural tube and constitute a ventral root. As early as the fifth week they are joined by sensory fibres of the dorsal root and continue as a nerve trunk.

Cells of the neural crest differentiate into things other than sensory neurons. Among these variants are cells that encapsulate ganglion cells and others that become neurolemma cells, which follow the peripherally growing nerve fibres and ensheath them. The neurolemma cells cover some nerve fibres with a fatty substance called myelin.

Spinal nerves. Spinal nerves are sensorimotor nerves, with dorsal and ventral roots. A network called a brachial plexus arises in relation to each upper limb and a lumbosacral plexus in relation to each lower limb. The spine, elongating faster than the spinal cord, drags nerve roots downward, since each nerve must continue to emerge between the same two vertebrae. Because of their appearance, the obliquely coursing nerve roots are named the cauda equina, the Latin term for horse's tail.

Cranial nerves. Cranial nerves V, VII, IX, and X arise in relation to embryonic branchial arches but have origins similar to the spinal nerves. The olfactory nerves (cranial nerve I) are unique in that their cell bodies lie in the olfactory epithelium (the surface membrane lining the upper parts of the nasal passages), each sending a nerve fibre back to the brain. The so-called optic nerves (II) are not true nerves but only tracts that connect the retina (a dislocated portion of the brain) with the brain proper. Nerves III, IV, VI, and XII are pure motor nerves that correspond to the ventral roots of spinal nerves. The acoustic nerves (VIII) are pure sensory nerves, each with a ganglion that subdivides for auditory functions and functions having to do with equilibrium and posture; they correspond to dorsal roots. Nerves X and XI are a composite of which XI is a motor component.

Autonomic nervous system. The autonomic nervous system is made up of two divisions, the sympathetic and the parasympathetic nervous systems; it controls such involuntary actions as constriction of blood vessels. Some cells of the neural crests migrate and form paired segmental masses alongside the aorta, a principal blood vessel. Part of the cells become efferent multipolar ganglion cells (cells whose fibres carry impulses outward from ganglions, or aggregates of nerve cells) and others merely encapsulate the ganglion cells. These autonomic ganglia link into longitudinal sympathetic trunks. Some of the neuroblasts migrate farther and assemble as collateral ganglia—ganglia not linked into longitudinal trunks. Still others migrate near, or within, the visceral organs that they will innervate and produce terminal ganglia. These ganglia are characteristic of the parasympathetic system.

Origins of autonomic nervous system

Some cells of certain primitive collateral ganglia leave and invade the amassing mesodermal cortex of each adrenal gland. Consolidating in the centre, they become the endocrine cells of the medulla.

Sense organs. *Olfactory organ.* Paired thickenings of ectoderm near the tip of the head infold and produce olfactory pits. These expand into sacs in which only a relatively small area becomes olfactory in function (Figure 30E). Some epithelial cells in these regions remain as inert supporting elements. Others become spindle-shaped olfactory cells. One end of each olfactory cell projects receptive olfactory hairs beyond the free surface of the epithelium. From the other end a nerve fibre grows back and makes a connection within the brain.

Gustatory organ. Most taste buds arise on the tongue. Each bud, a barrel-shaped specialization within the epithelium that clothes certain lingual papillae (small projections on the tongue), is a cluster of tall cells, some of which have differentiated into taste cells whose free ends bear receptive gustatory hairs. Sensory nerve fibres end at the surface of such cells. Other tall cells are presumably inertly supportive in function.

Eye. The earliest indication of an eye is an optic vesicle (sac) bulging from each side of the forebrain (Figure 30B, C). It quickly becomes an indented optic cup, connected to the brain by a slender optic stalk. Most of the cup will become the retina, but its rim represents the epithelial part of the insensitive ciliary body and iris. The thicker inner layer of the cup becomes the neural layer of the retina, and by the sixth month three strata of neurons are recognizable in it: (1) visual cells, each bearing either a photoreceptive rod or a cone at one end; (2) bipolar cells, intermediate in position; (3) ganglion cells, which sprout axons that grow back through the optic stalk and make connections within the brain. The thin outer layer of the cup remains a simple epithelium whose cells gain pigment and make up the pigment epithelium of the retina.

The lens arises as a thickening of the ectoderm adjacent to the optic cup. It inpockets to form a lens vesicle and then detaches. The cells of its back wall become tall, transparent lens fibres. Mesoderm surrounding the optic cup specializes into two accessory coats. The outer coat, the tough, white sclera, is continuous with the transparent cornea. The inner coat, the vascular choroid, continues as the vascular and muscular ciliary body and the vascularized tissue of the iris. The eyelids are folds of adjacent skin; from the inside of each upper lid several lacrimal glands bud out (see SENSORY RECEPTION: *Human vision: Structure and function of the eye*).

Ear. The projecting part (auricle) of the external ear develops from hillocks on the first and second branchial arches (Figure 30C–E). The ectodermal groove between those arches deepens and becomes the external auditory canal. The auditory tube and tympanic cavity—the cavity at the inner side of the eardrum—are expansions of the endodermal pouch located between the first and second branchial arches. The area where ectodermal groove and endodermal pouch come in contact is the site of the future eardrum. The chain of three auditory ossicles (small bones) that stretches across the tympanic cavity is a derivative of the first and second arches.

Origins of the ear

The epithelium of the internal ear is at first a thickening of ectoderm at a level midway of the hindbrain (Figure 30B). This plate inpockets and pinches off as a closed sac, the otocyst. Its ventral part elongates and coils to resemble a snail's shell, thereby forming the cochlear duct, or seat of the organ of hearing. A middle region of the otocyst becomes chambers known as the utricle and saccule, related to the sense of balance. The dorsal part of the otocyst remodels drastically into three semicircular ducts, related to the sense of rotation. Fibres of the acoustic nerve grow among specialized receptive cells differentiated in certain regions of these three divisions.

MESODERMAL DERIVATIVES

Skeletal system. Except for part of the skull, all bones pass through three stages of development: membranous, cartilaginous, and osseous. The earliest ossification centres appear in the eighth week, but some do not arise until childhood years and even into adolescence.

Axial skeleton. The ventromedial walls (the walls toward the front and the midline) of the paired somites break down, and their cells migrate toward the axial notochord and surround it. Differentiation and growth of these segmental masses produce the jointed vertebrae. Ribs also grow out of each primitive vertebral mass, but they become long only in the thoracic region. Here their ventral ends join the sternum, which arises independently by the fusion of a pair of bars.

The skull has three components, different in origin. Its

basal region is an ancient heritage whose bones pass through the three typical stages of development. By contrast, the sides and roof of the skull develop directly from membranous primordia, or rudiments. The jaws are derivatives of the first pair of cartilaginous branchial arches but develop as membrane bone. Ventral ends of the second to fifth arches contribute the cartilages of the larynx and the hyoid bone (a bone of horseshoe shape at the base of the tongue). Dorsal ends of the first and second arches become the three auditory ossicles (the small bones in the middle ear).

Appendicular skeleton. The limb bones develop in three stages from axial condensations in the local mesoderm. The shoulder and pelvic supports are comparable sets, as are the bones of the arms and legs.

Articulations. Some type of joint exists wherever bones meet. Joints that allow little or no movement consist of connective tissue, cartilage, or bone. Movable joints arise as fluid-filled clefts in mesoderm, which condenses peripherally inio a fibrous capsule.

Myoblast development

Muscular system. Much of each somite differentiates into myoblasts (primitive muscle cells) that become voluntary muscle fibres. Aggregations of such fibres become muscles of the neck and trunk. Muscles of the head and some of the neck muscles originate from mesoderm of branchial arches. Muscles of the limbs seemingly arise directly from local mesoderm. In general, muscle primordia may fuse into composites, split into subdivisions, or migrate away from their sites of origin. During these changes they retain their original nerve supply. Regardless of differences in source of origin, all voluntary muscle fibres are of the same striated type (marked by dark and light stripes). Spontaneous movements begin to occur in embryos about 10 weeks old. In general, involuntary muscle differentiates from mesoderm surrounding hollow organs; only the cardiac muscle type is striated.

Vascular system. All hollow organs, including arteries, veins, and lymphatics, are lined with epithelium—the principal functional tissue—and are ensheathed with muscular and fibrous coats.

Blood vessels. Primitive blood vessels arise in the mesoderm as tiny clefts bordered by flat endothelial cells. Growth and coalescence produce networks, out of which favoured channels persist as definite vessels, while others decline and disappear. A bilaterally symmetrical system of vessels is well represented in embryos four weeks old. This early plan is profoundly altered and made somewhat asymmetrical during the second month by fusions, atrophies, emergence of new vessels, and rerouting of older ones. The alterations reflect adjustments to changing form and pattern within the developing organ systems.

Arteries cranial to the heart (headward of the heart) are mostly products of the paired aortic arches, which course axially within the branchial arches, thus interconnecting the ventral aorta with paired dorsal aortas. The third pair of aortic arches becomes the common carotids; the fourth pair, the aortic arch and brachiocephalic; the fifth pair, the pulmonary arteries and ductus arteriosus. The dorsal aortas fuse into the single descending aorta, which bears three sets of paired, segmental branches. The dorsal set becomes the subclavian, intercostal, and lumbar arteries. The lateral set becomes arteries to the diaphragm, the adrenal glands, the kidneys, and the sex glands. The ventral set becomes the celiac, mesenteric, and umbilical arteries. Axial arteries to both sets of limb buds emerge from an original plexus, but they undergo drastic alteration and extensive replacement.

The primitive veins are symmetrically bilateral. They consist of vitelline veins from the yolk sac, umbilical veins from the placenta, and precardinal and postcardinal veins from the cranial and caudal regions (the regions toward the head and toward the tail) of the body. Drastic transformations occur in all of these, and new pairs of veins (subcardinals and supracardinals) arise also, caudal to the heart. From the vitellines come chiefly the portal and hepatic veins. The left umbilical becomes the main return from the placenta by making a diagonal channel, the ductus venosus, through the liver to the heart. The precardinal veins change their names to the internal jugulars, but near the heart an interconnection permits both to drain into a common stem, then called the superior vena cava. Caudal to the heart, the postcardinals virtually disappear, and all blood return shifts to the right side as a new compound vessel, the inferior vena cava, becomes dominant. Pulmonary veins open into the left atrium. Veins from the limb buds organize from an early peripheral border vein.

Lymphatic vessels. The lymph vessels develop independently in close association with veins. Linkages produce the thoracic duct, which is the main drainage return for lymph. Masses of lymphocytes accumulate about lymphatic vessels and organize as lymph nodes. The spleen has somewhat similar tissue, but its channels are supplied with blood.

Origins of the heart

Heart. Fusion combines two endothelial tubes, and these are surrounded by a mantle of mesoderm that will become the muscular and fibrous coats of the heart. At three weeks the heart is a straight tube that is beginning to beat (Figure 29M). Starting at the head end, four regions can be recognized: bulbus, ventricle, atrium, and sinus venosus. Since the heart is anchored at both ends, rapid elongation forces it to bend. In doing this, the sinus–atrium and bulbus–ventricle reverse their original relations. Further development concerns the transformation of a single-chambered heart into one with four chambers.

The atrium becomes subdivided by the growth of two incomplete partitions, or septa, placed close together and each covering the defect in the other. The ventricle also subdivides, but by a single, complete partition. A canal, connecting atria and ventricles, becomes two canals. The bulbus is absorbed into the right ventricle, and its continuation (the truncus) subdivides lengthwise, forming the aorta and the pulmonary artery. The right horn of the sinus venosus is absorbed into the right atrium, together with the superior and inferior venae cavae, which originally drained into the sinus. The transverse portion of the sinus persists as the coronary sinus. The pulmonary veins retain their early drainage into the left atrium. Important valves develop and ensure flow within the heart from atria to ventricles, and outward from the ventricles into the aorta and the pulmonary artery.

Birth initiates breathing, and the abandonment of the placental circulation follows. These changes entail a drastic rerouting of blood through the heart. As a result, the two atrial septa fuse and no longer permit blood to pass from the right atrium to the left atrium. Blood in the pulmonary artery no longer virtually bypasses the lungs; previously it had passed to the aorta directly through a shunt offered by the ductus arteriosus. As a sequel to these changes, the abandoned umbilical arteries, umbilical vein, ductus venosus, and ductus arteriosus all collapse and become fibrous cords.

Origins of the kidney, urinary bladder, and urethra

Urinary system. Vertebrates have made three experiments in kidney production: the pronephros, or earliest type; the mesonephros, or intermediate kidney; and the metanephros, or permanent kidney. All arise from the cellular plates called nephrotomes that connect somites with the mesodermal sheets that bound the body cavity. The vestigial pronephros is represented solely by several pairs of tubules; they join separately formed excretory ducts that grow downward and enter the cloaca, the common outlet for urine, genital products, and for intestinal wastes. Next tailward arise some 40 pairs of nephric (kidney) tubules that constitute the mesonephros; these tubules join the same excretory ducts, hereafter called the mesonephric ducts. The two sets of mesonephric tubules serve as functioning kidneys until the 10th week.

Each permanent kidney, or metanephros, develops still farther tailward. A so-called ureteric primordium buds off each mesonephric duct, near its hind end. The ureteric stem elongates and expands terminally, thereby forming the renal pelvis and calices; continued bushlike branching produces collecting ducts. The early ureteric bud invades a mass of nephrotome tissue. The branching collecting ducts progressively break this tissue up into tiny lumps, each of which becomes a long secretory tubule, or nephron, and joins a nearby terminal twig of the duct system. Continued proliferation of ducts and nephric tissue produces over a million urine-producing tubules in each kidney.

The blind caudal end of the endodermal hindgut absorbs the stem of each mesonephric duct, whereupon the remainder of the duct and the ureter acquire separate openings into the hindgut. This expanded region of the gut, now a potential receptacle for feces, urine, and reproductive products, is known as a cloaca. It next subdivides into a rectum behind and a urogenital sinus in front. The sinus, in turn, will specialize into the urinary bladder and the urethra. The prostate gland develops as multiple buds from the urethra, close to the bladder.

Origins of sex glands and ducts and external genitalia

Genital system. The genital organs begin to develop in the second month, but for a time sex is not grossly distinguishable. Also, a double set of male and female ducts arise, and not until later does the unneeded set decline. Hence, this period is commonly called the indifferent stage.

Gonads. Sex glands develop in a pair of longitudinal ridges located alongside the mesentery, the anchoring fold of membrane to the gut. The primordial sex cells appear first in the cloacal wall, from which they migrate upward in the gut, pass through its mesentery, and finally invade the genital ridges, where they proliferate. The testes are the earliest type of gonad to organize. They begin by developing testis cords and a testis capsule. The cords radiate from one focal point at the periphery, and thin fibrous partitions segregate groups of the cords within wedge-shaped compartments. These cords do not gain channels and become semen-producing tubules until near the time of puberty. The ovaries organize somewhat tardily by differentiating an outer portion, the cortex, and a central portion, the medulla. The cortex contains the primordial sex cells; these become surrounded by a layer of ordinary cells, thereby forming primary ovarian follicles. Both the testes and the ovaries undergo relative shifts from their early sites to lower positions in the body. But only the testes make a bodily descent; this is into the scrotum.

Genital ducts. In the male, a few mesonephric tubules on each side do not degenerate but link up with the neighbouring testis tubules. The converted mesonephric tubules and the retained mesonephric ducts become the male sex ducts. Near their terminations they outpouch seminal vesicles and then open into the urethra. In the female, a pair of ducts develops from the epithelium clothing the mesonephric ridges. These ducts, known as the uterine tubes, mostly parallel the courses of the mesonephric ducts, but at their lower ends they unite into a common tube that becomes the uterus and vagina.

External genitalia. Both sexes develop a genital tubercle (*i.e.*, a knob) and a pair of urogenital folds flanked by a pair of genital swellings. At three months these rudiments begin to assume male or female characteristics. In the male, the tubercle and the united urogenital folds combine as the penis, thereby continuing the urethra to its end; the genital swellings shift toward the anus, fuse, and become the scrotum. In the female, the tubercle remains small, as the clitoris; it does not contain the urethra. The urogenital folds remain unclosed as the lesser vulvar lips and are flanked by the unshifted and unfused genital swellings, or greater lips.

Origins of the body cavities

Coelom. The lateral mesoderm, beyond the somites and nephrotomes, splits into two layers: the somatic layer and, underlying the somatic layer, the splanchnic layer. The intervening space is the coelom. As the embryo's body folds off, its coelom becomes a single closed cavity. It can be recognized, regionally, a provisional pericardial cavity (cavity for the heart), two pleural canals (for the lungs), and a peritoneal cavity (for the abdominal contents). A thick plate of mesoderm, the transverse septum, constitutes a partial partition just ahead of the developing liver. Two pairs of membranes grow out from the septum. One set separates the pericardial cavity from the two pleural cavities; these membranes later expand into the pericardium and enclose the heart. The other pair of membranes separates the pleural cavities from the peritoneal cavity of the abdomen. The definitive diaphragm is a composite partition, much of which is furnished by the transverse septum; lesser contributions are from the lateral body walls and the paired membranes that separated the pleural and peritoneal cavities.

ENDODERMAL DERIVATIVES

Pharynx. The tongue is a product of four branchial arches, whose ventral ends merge in its midplane. Papillae elevate from the surface, and taste buds arise as specializations within the covering epithelium of some of them. Pharyngeal pouches are early lateral expansions of the local endoderm, alternating with the branchial arches. The first pair elongate as the auditory tubes and tympanic cavities. The second pair mark the site of the tonsils. The third pair give rise to the halves of the thymus, and the third and fourth pairs produce the two sets of parathyroid glands. The thyroid gland buds off the pharyngeal floor in the midplane and at the level of the second branchial arches.

Development of the esophagus, stomach, and intestines

Digestive tube. As the embryo folds off, the endoderm is rolled in as the foregut and hindgut. Continued growth progressively closes both the midbody and the midgut (Figure 29P, Q). The esophagus remains as a simple, straight tube. The stomach grows faster on its dorsal side, thereby forming the bulging greater curvature; the stomach also rotates 90° so that its original dorsal and ventral borders come to lie left and right. The intestine elongates faster than the trunk, so that its loops find temporary room by pushing into the umbilical cord. Later, the loops return, completing a rotation that gives the characteristic final placement of the small and large intestines.

When the gut folds into a tube it is suspended by a sheet-like dorsal mesentery, or membranous fold. In the region of the stomach it forms an expansive pouch, the omental bursa. Secondary fusions of the bursa and of some of the rest of the mesentery with the body wall produce lines of attachment from stomach to rectum inclusive, different from the original midplane course. Such fusions also anchor firmly some parts of the tract. A ventral mesentery, beneath the gut, exists only in the region of the stomach and liver.

Major glands. The liver arises as a ventral outgrowth of the foregut that invades the early transverse septum. Although rapid growth causes it to bulge prominently away from this septum, it remains attached to the septum and hence to the definitive diaphragm. The differentiating glandular tissue takes the form of plates bathed by blood channels. The stem of the original liver bud becomes the common bile duct, whereas a secondary outgrowth produces the cystic duct and the gallbladder.

The pancreas takes its origin from a larger dorsal bud and a smaller ventral bud, both off the foregut. The two merge and their ducts communicate, but in man it is the lesser, ventral duct that becomes the stem outlet. Secretory acini are berrylike endings of the branching ducts. Pancreatic islets arise as special sprouts from the ducts; these differentiate into endocrine tissue that secretes insulin.

Nose, larynx, trachea, and lungs

Respiratory system. *Nasal cavity.* The first part of the respiratory system is ectodermal in origin. The olfactory sacs (see above *Ectodermal derivatives: Sense organs*) become continuous secondarily with a passage captured from the primitive mouth cavity. This addition is produced by a horizontal partition, the palate. It arises from a pair of shelflike folds that grow out from the halves of the primitive upper jaw and then unite. The final nasal passage extends from the nostrils to the back of the pharynx.

Larynx, trachea, and lungs. A hollow lung bud grows off the floor of the endodermal pharynx, just caudal (tailward) to the pharyngeal pouches and in the midline. It has the form of a tube with an expanded end. The entrance to this tube is the glottis, and the region about it becomes the larynx. The tube proper represents the trachea. Its terminal expansion divides into two branches, and these tubes elongate as the primary bronchi. Continued growth and budding produce two side branches from the right bronchus and one from the left. These branches and the blind ends of the two parent bronchi indicate the future plan of the lungs, with three right lobes and two left lobes. Continued branchings, through the sixth month, produce bronchioles of different orders. In the final months the smaller ducts and early respiratory alveoli (air sacs) appear, the lungs losing their previous glandular appearance and also becoming highly vascular. Until breathing distends the lungs, these organs remain relatively small.

(L.B.A.)

Postnatal development

Human growth is far from being a simple and uniform process of becoming taller or larger. As a child gets bigger, there are changes in shape and in tissue composition and distribution. In the newborn infant the head represents about a quarter of the total length; in the adult it represents about one-seventh. In the newborn infant the muscles constitute a much smaller percentage of the total body mass than in the young adult. In most tissues, growth consists both of the formation of new cells and the packing in of more protein or other material into cells already present; early in development cell division predominates and later cell filling.

TYPES AND RATES OF HUMAN GROWTH

The complex patterns of growth

Different tissues and different regions of the body mature at different rates, and the growth and development of a child consists of a highly complex series of changes. It is like the weaving of a cloth whose pattern never repeats itself. The underlying threads, each coming off its reel at its own rhythm, interact with one another continuously, in a manner always highly regulated and controlled. The fundamental questions of growth relate to these processes of regulation, to the program that controls the loom, a subject as yet little understood. Meanwhile, height is in most circumstances the best single index of growth, being a measure of a single tissue (that of the skeleton; weight is a mixture of all tissues, and this makes it a less useful parameter in a long-term following of a child's growth). In this section, the height curves of girls and boys are considered in the three chief phases of growth; that is (briefly) from conception to birth, from birth until puberty, and during puberty. Also described are the ways in which other organs and tissues, such as fat, lymphoid tissue, and the brain, differ from height in their growth curves. There is a brief discussion of some of the problems that beset the investigator in gathering and analyzing data about growth of children, of the genetic and environmental factors that affect rate of growth and final size, and of the way hormones act at the various phases of the growth process. Lastly, there is a brief look at disorders of growth. Throughout, the emphasis is on ways in which individuals differ in their rates of growth and development.

Growth as distance and velocity

The changes in height of the developing child can be thought of in two different ways: the height attained at successive ages and the increments in height from one age to the next, expressed as rate of growth per year. If growth is thought of as a form of motion, the height attained at successive ages can be considered the distance travelled, and the rate of growth, the velocity. The velocity or rate of growth reflects the child's state at any particular time better than does the height attained, which depends largely on how much the child has grown in all preceding years. The blood and tissue concentrations of those substances whose amounts change with age are thus more likely to run parallel to the velocity rather than to the distance curve. In some circumstances, indeed, it is the acceleration rather than the velocity curve that best reflects physiological events.

In general, the velocity of growth decreases from birth onward (and actually from as early as the fourth month of fetal life; see below), but this decrease is interrupted shortly before the end of the growth period. At this time, in boys from about 13 to 15 years (Figure 33), there is marked acceleration of growth, called the adolescent growth spurt. From birth until age four or five, the rate of growth in height declines rapidly, and then the decline, or deceleration, gets gradually less, so that in some children the velocity is practically constant from five or six up to the beginning of the adolescent spurt. A slight increase in velocity is sometimes said to occur between about six and eight years.

This general velocity curve of growth in height begins a considerable time before birth. Figure 31 shows the distance and velocity curves for body length in the prenatal period and first two postnatal years. The peak velocity of length is reached at about four months after the mother's last menstruation. (Age in the fetal period is usually reckoned from the first day of the last menstrual period, an average of two weeks before actual fertilization, but, as a rule, the only locatable landmark.)

Growth in weight of the fetus follows the same general pattern as growth in length, except that the peak velocity is reached much later, at approximately 34 weeks after the mother's last menstrual period.

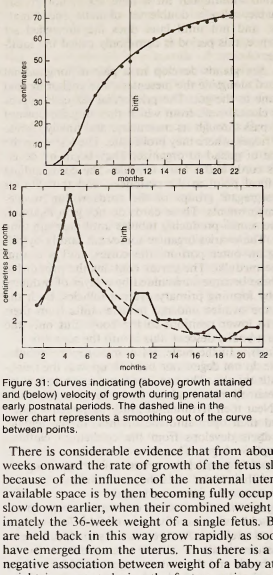


Figure 31: Curves indicating (above) growth attained and (below) velocity of growth during prenatal and early postnatal periods. The dashed line in the lower chart represents a smoothing out of the curve between points.

There is considerable evidence that from about 34 to 36 weeks onward the rate of growth of the fetus slows down because of the influence of the maternal uterus, whose available space is by then becoming fully occupied. Twins slow down earlier, when their combined weight is approximately the 36-week weight of a single fetus. Babies who are held back in this way grow rapidly as soon as they have emerged from the uterus. Thus there is a significant negative association between weight of a baby at birth and weight increment during the first year; in general, larger babies grow less, the smaller more. For the same reason there is practically no relation between adult size and the size of that person at birth, but a considerable relation has developed by the time the person is two years old. This slowing-down mechanism enables a genetically large child developing in the uterus of a small mother to be delivered successfully. It operates in many species of animals; the most dramatic demonstration was by crossing reciprocally a large Shire horse and a small Shetland pony. The pair in which the mother was a Shire had a large newborn foal, and the pair in which the mother was Shetland had a small foal. But both foals were the same size after a few months, and when fully grown both were about halfway between their parents. The same has been shown in cattle crosses.

Rapid growth following small size at birth

Poor environmental circumstances, especially of nutrition, result in lowered birth weight in the human being. This seems chiefly to be caused by a reduced rate of growth in the last two to four weeks of fetal life, for weights of babies born in 36 or 38 weeks in various parts of the world in various circumstances are said to be similar. Mothers who, because of adverse circumstances in their own childhood, have not achieved their full growth potential may produce smaller fetuses than they would have, had they grown up in better circumstances. Thus two generations or even more may be needed to undo the effect of poor environmental circumstances on birth weight.

The great rate of growth of the fetus compared with that of the child is largely due to the fact that cells are still multiplying. The proportion of cells undergoing mitosis (the ordinary process of cell multiplication by splitting) in any tissue becomes progressively less as the fetus gets older, and it is generally thought that few if any new nerve cells (apart from the cells in the supporting tissue, or neuroglia) and only a limited proportion of new muscle cells appear after six postmenstrual months, the time when the velocity in linear dimensions is dropping sharply.

Prenatal and postnatal growth contrasted

The muscle and nerve cells of the fetus are considerably different in appearance from those of the child or adult. Both have little cytoplasm (cell substance) around the nucleus. In the muscle there is a great amount of intercellular substance and a much higher proportion of water than in mature muscle. The later fetal and the postnatal growth of the muscle consists chiefly of building up the cytoplasm of the muscle cells; salts are incorporated and the contractile proteins formed. The cells become bigger, the intercellular substance largely disappears, and the concentration of water decreases. This process continues quite actively up to about three years of age and slowly thereafter; at adolescence it briefly speeds up again, particularly in boys, under the influence of androgenic (male sex) hormones. In the nerve cells cytoplasm is added and elaborated, and extensions grow that carry impulses from and to the cells—the axons and dendrites, respectively. Thus postnatal growth, for at least some tissues, is chiefly a period of development and enlargement of existing cells, while early fetal life is a period of division and addition of new cells.

TYPES OF GROWTH DATA

Growth is in general a regular process. Contrary to what is said in some of the older textbooks, growth in height does not proceed by fits and starts, nor does growth in upward dimensions alternate with growth in transverse ones. The more carefully measurements are taken, with precautions, for example, to minimize the decrease in height that occurs during the day for postural reasons, the more regular does the succession of points in a graph of growth become. Many attempts have been made at finding mathematical curves that fit, and thus summarize, human growth data. What is needed is a curve or curves with relatively few constants, each capable of being interpreted in a biologically meaningful way. Yet the fit to empirical data must be adequate within the limits of measuring error. The problem is difficult, partly because the measurements usually taken are themselves biologically complex. Stature, for example, consists of leg length and trunk length and head height, all of which have rather different growth curves. Even with relatively homogeneous dimensions such as the length of the radius bone in the forearm, or width of an arm muscle, it is not clear what purely biological assumptions should be made as the basis for the form of the curve. The assumption that cells are continuously dividing leads to a different formulation from the assumption that cells are adding constant amounts of nondividing material or amounts of nondividing material at rates varying from one age period to another.

Fitting a curve to the individual values, however, is the only way of extracting the maximum information from an individual's measurement data. More than one curve is needed to fit the postnatal age range. It seems that two curves may suffice, at least for many measurements such as height and weight—one curve for the period from a few months after birth to the beginning of adolescence and a different type of curve for the adolescent spurt.

Advantages in longitudinal and cross-sectional studies

Such curves have to be fitted to data on single individuals. Yearly averages derived from different children each measured only once do not, in general, give the same curve. Thus the distinction between the two sorts of investigation is important. When the same child at each age is used, the study is called longitudinal; when different children at each age are used, it is called cross-sectional. In a cross-sectional study all of the children at age eight, for example, are different from those at age seven. A study may be longitudinal over any number of years; there are short-term longitudinal studies extending from age four to six, for instance, and full birth-to-maturity longitudinal studies in which the children may be examined once, twice, or more times every year from birth until 20 or over. Mixed longitudinal studies are those in which children join and leave the group studied at varying intervals. Both cross-sectional and longitudinal studies have their uses, but they do not give the same information, and the same statistical methods cannot be used for the two types of study. Cross-sectional surveys are obviously cheaper and more quickly done and can include much larger numbers of children. Periodic cross-sectional surveys are valuable in assessing the nutritional progress of a country or a socioeconomic group and the health of the child population as a whole. But they never reveal individual differences in rate of growth or in the timing of particular phases such as the adolescent growth spurt. It is these individual rate differences that throw light on the genetic control of growth and on the correlation of growth with psychological development, educational achievement, and social behaviour.

Longitudinal studies are laborious and time-consuming; they demand great perseverance on the part of those who make them and those who take part in them; and they demand high technical standards, since in the calculation of a growth increment from one occasion to the next opportunities for two errors of measurement occur. In spite of these problems, longitudinal studies are the indispensable base on which the diagnosis and treatment of disorders of growth rest, for the clinical approach is a longitudinal one; and each child treated with human growth hormone, or with other hormones that affect growth, represents an attempt to alter an individual pattern of growth velocity.

Averages simply computed from cross-sectional data inevitably produce velocity curves that are flatter and broader than the curve for an individual and hence not a proper basis for clinical standards. It is possible to construct curves, however, whose 50th percentile (or average) represents the actual growth of a typical individual, by taking the shape of the curve from individual longitudinal data and the absolute values for the beginning and end from large cross-sectional surveys. Figures 32 and 33 show height-attained and height-velocity curves for the "typical" boy and girl in Britain in 1965, determined in this way. By "typical" is meant that boy or girl who has the mean (average) birth length, grows always at the mean velocity, has the peak of the adolescent growth spurt at the mean age, and finally reaches the mean adult height at the mean age of cessation of growth. Practically no individual follows the 50th percentile curve of Figure 32, but most have curves of the same shape. Standards for height for clinical use are constructed around these curves.

From *Archives of Disease in Childhood* (1966)

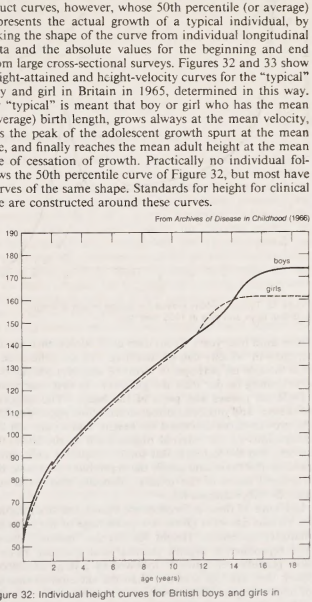


Figure 32: Individual height curves for British boys and girls in 1965 (see text).

Measuring the child's height

Boys' and girls' height curves. Figures 32 and 33 show the height curves from birth to maturity. Up to age two, the child is measured lying on his back. One examiner holds his head in contact with a fixed board, and a second person stretches him out to his maximum length and then brings a moving board into contact with his heels. This measurement, called supine length, averages about one centimetre more than the measurement of standing height taken on the same child, hence the break in the line in Figure 32 at age two. This occurs even when, as in the best techniques, the child is urged to stretch upwards to the full and is aided in doing so by a measurer's applying gentle upward pressure to his mastoid processes.

Figure 32 shows the typical girl as slightly shorter than the typical boy at all ages until adolescence. She becomes taller shortly after age 11 because her adolescent spurt takes place two years earlier than the boy's. At age 14 she is surpassed again in height by the typical boy, whose adolescent spurt has now started, while hers is nearly finished. In the same way, the typical girl weighs a little less than the boy at birth, equals him at age eight, becomes heavier at age nine or 10, and remains so until about age 14½.

The velocity curves given in Figure 33 show these processes more clearly. At birth the typical boy is growing slightly faster than the typical girl, but the velocities become equal at about seven months, and then the girl grows

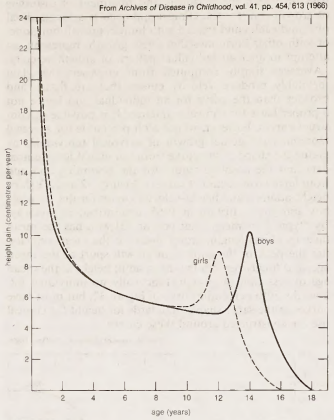


Figure 33: Typical velocity curves for supine length or height in British boys and girls in 1965 (see text).

faster until four years. From then until adolescence no differences in velocity can be detected. The sex difference is best thought of, perhaps, in terms of acceleration, the boy decelerating harder than the girl over the first four years.

Different tissues and parts of the body. The majority of skeletal and muscular dimensions follow approximately the growth curve described for height, and so also do the dimensions of the internal organs such as the liver, the spleen, and the kidneys. But some exceptions exist, most notably the brain and skull, the reproductive organs, the lymphoid tissue of the tonsils, adenoids, and intestines, and the subcutaneous fat.

In Figure 34 these differences are shown; the size attained by various tissues is given as a percentage of the birth-to-maturity increment. Height follows the "general" curve. The reproductive organs, internal and external, have a slow prepubescent growth, followed by a large adolescent spurt; they are less sensitive than the skeleton to one set of hormones and more sensitive to another.

The brain, together with the skull covering it and the eyes and ears, develops earlier than any other part of the body and thus has a characteristic postnatal curve. At birth it is already 25 percent of its adult weight, at age five about 90 percent, and at age 10 about 95 percent. Thus if the brain has any adolescent spurt at all, it is a small one. A small but definite spurt occurs in head length and breadth, but all or most of this is due to thickening of the skull bones and the scalp, together with development of the air sinuses.

Growth of the brain, the skull, and the face

The dimensions of the face follow a path somewhat closer to the general curve. There is a considerable adolescent spurt, especially in the lower jaw, or mandible, resulting in the jaw's becoming longer and more projecting, the profile straighter, and the chin more pointed. As always

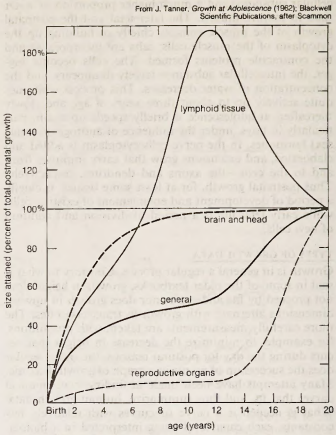


Figure 34: Major types of postnatal growth of parts and organs of the body (see text).

in growth, there are considerable individual differences, to the point that a few children have no detectable spurt at all in some face measurements.

The eye probably has a slight adolescent spurt, which is probably responsible for the increase in frequency of short-sightedness in children that occurs at the time of puberty. Though the degree of myopia increases continuously from at least age six to maturity, a particularly rapid rate of

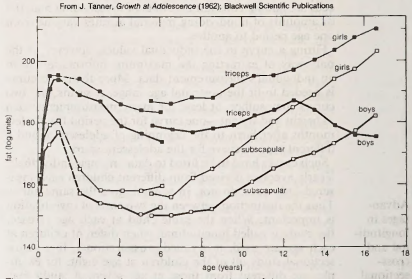


Figure 35: Amount of subcutaneous fat on the back of the arm (triceps) and the back of the body (subscapular) from birth to 16 years (see text).

change occurs at about 11 to 12 in girls and 13 to 14 in boys, and this would be expected if there was a rather greater spurt in the axial dimension (the dimension from front to back) of the eye than in its vertical dimension.

The lymphoid tissue has quite a different growth curve from the rest. It reaches its maximum amount before adolescence and then, probably under the direct influence of sex hormones, declines to its adult value.

Development of fat

The subcutaneous fat layer also has a curve of its own, of a slightly complicated sort. Its thickness can be measured either by X-rays or, more simply, at certain sites in the body, by picking up a fold of skin and fat between the thumb and forefinger and measuring its thickness with a special, constant-pressure caliper. Figure 35 shows the distance curves of skin folds taken halfway down the back of the upper arm (the triceps) and on the body just below the angle of the shoulder blade, or scapula. Subcutaneous fat begins to be laid down in the fetus at about 34 weeks postmenstrual age, increases from then until birth and from birth onward until about nine months. (This is in the average child; the peak may be reached as early as six months or as late as 12 or 15.) After nine months, when the velocity of fat gain is zero, the fat usually decreases (that is, it has a negative velocity) until age six to eight, when it begins to increase once more. Girls have a little more fat than boys at birth, and the difference becomes more marked during the period of loss, since girls lose less than boys. From eight years on, the curves for girls and boys diverge more radically, as do the curves for limb and body fat. At adolescence the limb fat in boys decreases, while the body fat shows a temporary slowing down of gain but no actual loss. In girls there is a slight halting of the limb-fat gain at adolescence, but no loss; the trunk fat shows only a steady rise until adolescence.

Development at puberty

ALTERATIONS IN GROWTH RATE

At puberty, a considerable alteration in growth rate occurs. There is a swift increase in body size, a change in shape and composition of the body, and a rapid development of the gonads, or sex glands—the reproductive organs and the characters signalling sexual maturity. Some of these changes are common to both sexes, but most are sex-specific. Boys have a great increase in muscle size and strength, together with a series of physiological changes making them capable of doing heavier physical work than girls and of running faster and longer. These changes all specifically adapt the male to his primitive primate role of dominating, fighting, and foraging. Such adolescent changes occur generally in primates (that is, men, apes, and monkeys) but are more marked in some species than in others. Man lies at about the middle of the primate range, as regards both adolescent size increase and degree of sexual differentiation.

During the adolescent spurt in height, for a year or more, the velocity of growth approximately doubles; a boy is likely to be growing again at the rate he last experienced about age two. The peak velocity of height (P.H.V., a point much used in growth studies) averages about 10.5 centimetres per year in boys and 9.0 centimetres in girls (about 4 and 3.4 inches, respectively), but this is the "instantaneous" peak given by a smooth curve drawn through the observations. The velocity over the whole year encompassing the six months before and after the peak is naturally somewhat less. During this year a boy usually grows between 7 and 12 centimetres (2.75 and 4.75 inches) and a girl between 6 and 11 centimetres (2.35 and 4.35 inches). Children who have their peak early reach a somewhat higher peak than those who have it late.

Age of peak height

The average age at which the peak is reached depends on the nature and circumstances of the group studied more, probably, than does the height of the peak. In moderately well-off British or North American children at present the peak occurs on average at about 14.0 years in boys and 12.0 years in girls. Though the absolute average ages differ from population to population, the two-year sex difference always persists.

Practically all skeletal and muscular dimensions take part in the spurt, though not to an equal degree. Most of the spurt in height is due to acceleration of trunk length rather than of length of legs. There is a fairly regular order in which the dimensions accelerate; leg length as a rule reaches its peak first, followed by the body breadths, with shoulder width last. The earliest structures to reach their adult status are the head, hands, and feet.

The spurt in muscle, of both limbs and heart, coincides with the spurt in skeletal growth, for both are caused by the same hormones. Boys' muscle widths reach a peak velocity of growth that is greater than that reached by girls. But, since girls have their spurt earlier, there is actually a period, from about 12½ to 13½, when girls on average have larger muscles than boys of the same age, as well as being taller (Figure 32). Simultaneously with the spurt there is a loss of fat, as described above.

The marked increase in muscle size in boys at adolescence leads to an increase in strength, illustrated in Figure 36. Before adolescence, boys and girls are similar in strength for a given body size and shape; after, boys

From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications

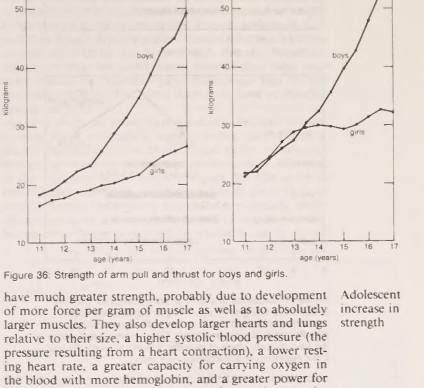


Figure 36. Strength of arm pull and thrust for boys and girls.

have much greater strength, probably due to development of more force per gram of muscle as well as to absolutely larger muscles. They also develop larger hearts and lungs relative to their size, a higher systolic blood pressure (the pressure resulting from a heart contraction), a lower resting heart rate, a greater capacity for carrying oxygen in the blood with more hemoglobin, and a greater power for neutralizing the chemical products of muscular exercise such as lactic acid. In short, the male becomes at adolescence more adapted for the tasks of hunting, fighting, and manipulating all sorts of heavy objects, as is necessary in some forms of food gathering.

Adolescent increase in strength

It is as a direct result of these anatomical and physiological changes that athletic ability increases so much in boys at adolescence. The popular notion of a boy's "outgrowing his strength" at this time has little scientific support. It is true that the peak velocity of strength is reached a year or so later than that of height, so that a short period may exist when the adolescent, having completed his skeletal and probably also his muscular growth, still does not have the strength of a young adult of the same body size and shape. But this is a temporary phase; considered absolutely, power, athletic skill, and physical endurance all increase progressively and rapidly throughout adolescence.

The adolescent spurt in skeletal and muscular dimensions is closely related to the rapid development of the reproductive system that takes place at this time. The course of this development is outlined diagrammatically in Figure 37. The bar marked "breast" in the chart for the girls and the bars marked "penis" and "testis" in the chart for the boys represent the periods of accelerated growth of these organs. Other bars indicate the genitalia rating and the advent and development of the pubic hair. The sequence

Development of reproductive organs

and timings that are given represent in each case average values for British boys and girls; the North American average is within two or three months of this. To give an idea of the individual departures from the average, figures for the range of age at which the various events begin and end are inserted under the first and last point of the bars. The acceleration of penis growth, for example, begins on average at about age 12½ years, but sometimes as early as 10½ and sometimes as late as 14½. The completion of penis development usually occurs at about age 14½, but in some boys is at 12½ and in others at 16½. There are a few boys, it will be noticed, who do not begin their spurts in height or penis development until the earliest

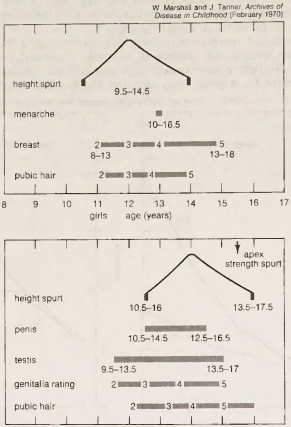


Figure 37. Sequence of events at adolescence for average British boys and girls. Range of ages within which each event may begin and end is given by the figures placed directly below its start and finish. The numbers 2 to 5 within the bars indicate stage of maturity, 5 being full maturity and 2 the beginning of adolescence (see text).

maturers have entirely completed theirs. At ages 13, 14, and 15 there is an enormous variability among any group of boys, who range all the way from practically complete maturity to absolute preadolescence. The same is true of girls aged 11, 12, and 13.

The psychological and social importance of this difference in the tempo of development, as it has been called, is great, particularly in boys. Boys who are advanced in development are likely to dominate their contemporaries in athletic achievement and sexual interest alike. Conversely the late developer is the one who all too often loses out in the rough and tumble of the adolescent world, and he may begin to wonder whether he will ever develop his body properly or be as well endowed sexually as those others whom he has seen developing around him. An important part of the educationist's and the doctor's task at this time is to provide information about growth and its variability to preadolescents and adolescents and to give sympathetic support and reassurance to those who need it.

The first sign of puberty in the boy

The sequence of events, though not exactly the same for each boy or girl, is much less variable than the age at which the events occur. The first sign of puberty in the boy is usually an acceleration of the growth of the testes and scrotum with reddening and wrinkling of the scrotal skin. Slight growth of pubic hair may begin about the same time but is usually a trifle later. The spurts in height and penis growth begin on average about a year after the first testicular acceleration. Concomitantly with the growth of the penis, and under the same stimulus, the seminal vesicles, the prostate, and the bulbo-urethral glands, all of which contribute their secretions to the seminal fluid, enlarge and develop. The time of the first ejaculation of seminal fluid is to some extent culturally as well as biologically determined but as a rule is during adolescence and about a year after the beginning of accelerated penis growth.

Axillary (armpit) hair appears on average some two years after the beginning of pubic hair growth; that is, when pubic hair is reaching stage 4 (Figure 37). There is enough variability and dissociation in these events, so that a very few children's axillary hair actually appears first. In boys, facial hair begins to grow at about the time that the axillary hair appears. There is a definite order in which the hairs of moustache and beard appear: first at the corners of the upper lip, then over all the upper lip, then at the upper part of the cheeks, in the midline below the lower lip, and, finally, along the sides and lower borders of the chin. The remainder of the body hair appears from about the time of first axillary hair development until a considerable time after puberty. The ultimate amount of body hair that an individual develops seems to depend largely on heredity, though whether because of the kinds and amounts of hormones secreted or because of variations in the reactivity of the end organs is not known.

Breaking of the voice occurs relatively late in adolescence. The change in pitch accompanies enlargement of the larynx and lengthening of the vocal cords, caused by the action of the male hormone testosterone on the laryngeal cartilages. There is also a change in quality that distinguishes the voice (more particularly the vowel sounds) of both male and female adults from that of children. This is caused by the enlargement of the resonating spaces above the larynx, as a result of the rapid growth of the mouth, nose, and maxilla (upper jaw).

In the skin, particularly of the armpits and the genital and anal regions, the sebaceous and apocrine sweat glands develop rapidly during puberty and give rise to a characteristic odour; the changes occur in both sexes but are more marked in the male. Enlargement of the pores at the root of the nose and the appearance of comedones (blackheads) and acne, while likely to occur in either sex, are considerably more common in adolescent boys than girls, since the underlying skin changes are the result of androgenic (male sex hormone) activity.

During adolescence the male breast undergoes changes, some temporary and some permanent. The diameter of the areola, which is equal in both sexes before puberty, increases considerably, though less than it does in girls. In some boys (between a fifth and a third of most groups studied) there is a distinct enlargement of the breast (sometimes unilaterally) about midway through adolescence. This usually regresses again after about one year.

First signs of puberty in the girl

In girls the start of breast enlargement—the appearance of the "breast bud"—is as a rule the first sign of puberty, though the appearance of pubic hair precedes it in about one-third. The uterus and vagina develop simultaneously with the breast. The labia and clitoris also enlarge. Menarche, the first menstrual period, is a late event in the sequence. Though it marks a definitive and probably mature stage of uterine development, it does not usually signify the attainment of full reproductive function. The early cycles may be more irregular than later ones and in some girls, but by no means all, are accompanied by discomfort. They are often anovulatory; that is, without the shedding of an egg. Thus there is frequently a period of adolescent sterility lasting a year to 18 months after menarche, but it cannot be relied on in the individual case. Similar considerations may apply to the male, but there is no reliable information about this. On average, girls grow about six centimetres (about 2.4 inches) more after menarche, though gains of up to twice this amount may occur. The gain is practically independent of whether menarche occurs early or late.

NORMAL VARIATIONS

Children vary a great deal both in the rapidity with which they pass through the various stages of puberty and in

Stages of puberty

the closeness with which the various events are linked together. At one extreme one may find a perfectly healthy girl who has not yet menstruated though her breasts and pubic hair are characteristic of the adult and she is already two years past her peak height velocity; and at the other, a girl who has passed all the stages of puberty within the space of two years.

In girls the interval from the first indication of puberty to complete maturity varies from 18 months to six years. The period from the moment when the breast bud first appears to menarche averages 2½ years, but it may be as little as six months or as much as 5½ years. The rapidity with which a child passes through puberty seems to be independent of whether puberty is occurring early or late. Menarche invariably occurs after peak height velocity has been passed.

In boys a similar variability of maturation occurs. The male genitalia may take between two and five years to attain full development, and some boys complete the whole process before others have moved from the first to the second stage.

The height spurt occurs relatively later in boys than in girls. Thus there is a difference between the average boy and girl of two years in age of peak height velocity but of only one year in the first appearance of pubic hair. Indeed, in some girls the acceleration in height is the first sign of puberty; this is never so in boys. A small boy whose genitalia are just beginning to develop can be unequivocally reassured that an acceleration in height is soon to take place, but a girl in the corresponding situation may already have had her height spurt.

Sex dimorphism. The differential effects on the growth of bone, muscle, and fat at puberty increase considerably the difference in body composition between the sexes. Boys have a greater increase not only in stature but especially in breadth of shoulders; girls have a greater relative increase in width of hips. These differences are produced chiefly by the changes that occur during puberty, but other sex differentiations arise before that time. Some, like the external genital difference itself, develop during fetal life. Others develop continuously throughout the whole growth period by a sustained differential growth rate. An example of this is the greater relative length and breadth of the forearm in the male when compared with whole arm length or whole body length.

Part of the sex difference in pelvic shape antedates puberty. Girls at birth already have a wider pelvic outlet. Thus the adaptation for childbearing is present from an early age. The changes at puberty are concerned more with widening the pelvic inlet and broadening the much more noticeable hips.

Physical and behavioral interaction. Children vary greatly in their tempo of growth. The effects are most dramatically seen at adolescence, but they are present at all ages from birth and even before.

Developmental and chronological ages

The concept of developmental age, as opposed to chronological age, is an important one. To measure developmental age, there is need of some way of determining how far along his own path to maturity a given child has gone. Therefore, there is need of a measure in which everyone at maturity ends up the same (not different as in height). The usual measure used is skeletal maturity or bone age. This is measured by taking an X-ray of the hand and wrist. The appearances of the developing bones can be rated and formed into a scale of development; the scale is applicable to boys and girls of all genetic backgrounds, though girls on average reach any given score at a younger age than do boys; and blacks on average, at least in the first few years after birth, reach a given score younger than do whites. Other areas of the body may be used if required. Skeletal maturity is closely related to the age at which adolescence occurs; that is, to maturity measured by some sex character developments. Thus the range of the chronological age within which menarche may normally fall is about 10 to 16½, but the corresponding range of bone age for menarche is only 12 to 14½. Evidently the physiological processes controlling progression of skeletal development are in most instances closely linked with those that initiate the events of adolescence. Furthermore, children tend to be consistently advanced or retarded during their whole growth period, at any rate after about age three.

There is little doubt that being an early or a late maturer may have repercussions on behaviour and that in some children these repercussions may be considerable. There is little enough solid information on the relation between emotional and physiological development, but what there is supports the common-sense notion that emotional attitudes are clearly related to physiological events.

Effects of late and early maturity

Larger size and earlier maturation. The rate of maturing and the age of onset of puberty are dependent on a complex interaction of genetic and environmental factors. Where the environment is good, most of the variability in age at menarche in a population is due to genetical differences. In many societies puberty occurs later in the poorly off, and, in most societies investigated, children with many siblings grow more slowly than children with few.

Modern tendency toward earlier maturity

During the last hundred years there has been a striking tendency for children to become progressively larger at all ages. This is known as the "secular trend." The magnitude of the trend in Europe and America is such that it dwarfs the differences between socioeconomic classes.

The data from Europe and America agree well: from about 1900, or a little earlier, to the present, children in average economic circumstances have increased in height at age five to seven by about one to two centimetres (0.4 to 0.8 inch) per decade, and at 10 to 14 by two to three centimetres (0.8 to 1.2 inches) each decade. Preschool data show that the trend starts directly after birth and may, indeed, be relatively greater from age two to five than subsequently. The trend started, at least in Britain, as early as 1850.

Most of the trend toward greater size in children reflects a more rapid maturation; only a minor part reflects a greater ultimate size. The trend toward earlier maturing is best shown in the statistics on age at menarche. A selection of the best data is illustrated in Figure 38. The trend is between three and four months per decade since 1850 in average sections of western European populations. Well-off persons show a trend of about half of this magnitude, having never been so retarded in menarche as the worse

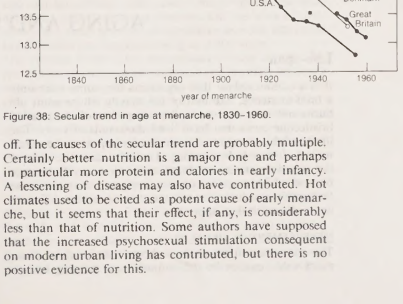


Figure 38. Secular trend in age at menarche, 1830–1960.

off. The causes of the secular trend are probably multiple. Certainly better nutrition is a major one and perhaps in particular more protein and calories in early infancy. A lessening of disease may also have contributed. Hot climates used to be cited as a potent cause of early menarche, but it seems that their effect, if any, is considerably less than that of nutrition. Some authors have supposed that the increased psychosexual stimulation consequent on modern urban living has contributed, but there is no positive evidence for this.

Hormones and growth

The main hormones concerned with growth are pituitary growth hormone, thyroid hormone, the sex hormones testosterone and estrogen, and the pituitary gonadotropic (sex-gland-stimulating) hormones.

Action of growth hormone

Pituitary growth hormone, a protein with molecular weight of 21,600 and of known amino-acid composition, is secreted by the pituitary gland throughout life. Exactly what its function is in the adult is not clear, but in the child it is necessary for growth; without it dwarfism results. During fetal life it seems not to be necessary, though normally present. It is not secreted at a constant rate all day but in small bursts of activity. Secretion by the pituitary is controlled by a substance sent to it from an adjacent part of the brain. The normal stimulus for secretion is not certain, but a sharp and "unnatural" lowering of blood sugar will cause growth hormone to be secreted, and this is used as a test. The hormone decreases the amount of fat and causes protein to be laid down in muscles and viscera. Children who lack it are fat as well as small; when given it by injection, they lose fat and grow rapidly.

The hormone is peculiar in being species-specific; that is, only growth hormone from human glands is active in man. Supplies of the hormone for treating children who need it are obtained at autopsy, and supply has been limited by this. Recombinant DNA technology shows possibilities in increased manufacture of this hormone in the laboratory.

Thyroid hormone from the thyroid gland in the neck is necessary for normal growth, though it does not itself stimulate growth, for example, in the absence of pituitary growth hormone. Without thyroid hormone, however, cells do not develop and function properly, especially in the brain. Babies who lack thyroid hormone at birth are small and have insufficiently developed brains; they are known as cretins. Frequently, if the condition is diagnosed and they are treated with thyroid hormone at once, they recover completely; the longer they go without treatment, the more likely it is that the brain damage will be permanent.

Thyroid lack may also develop later in childhood, when it causes a slowing of growth rate; full catch-up follows prompt treatment.

Testosterone, secreted by the interstitial cells of the testis, is important not only at puberty but before. Its secretion by the fetal testis cells is responsible for the development of certain parts of the male genital apparatus. If testosterone is not secreted at a particular and circumscribed time, the genitalia develop into the female form.

Effects of sex hormones on growth

Only small amounts of testosterone circulate between birth and puberty, but at puberty the interstitial cells develop greatly in response to pituitary luteinizing hormone (see below), and testosterone is secreted in large amounts, bringing about most of the changes of male puberty. It acts on a widespread series of receptors—for example, the cells of the penis, the muscles, the skin of the face, the cartilages of the shoulder, and certain parts of the brain. Most of the adolescent growth spurt is due to testosterone.

The female sex hormones, collectively called estrogens, are first secreted in quantity at puberty by cells in the ovary. They cause growth of the uterus, vagina, and breast; they act also on the bones of the hip, causing the specifically female widening. The adolescent growth spurt in the female is probably caused by testosterone-like substances (androgens) secreted by the adrenal gland in both male and female.

Initiation of puberty

The pituitary secretes two other hormones concerned in development: one, follicle-stimulating hormone (FSH), causes growth of the main portions of the ovary in the female and the sperm-producing cells in the testis of the male; the other, luteinizing hormone (LH), causes growth and secretion of the testosterone-secreting cells of the male and has an action in controlling the menstrual cycle in the female. The pituitary is caused to secrete gonadotropins by substances called releasing factors that come to it from adjacent areas of the brain, where they are made. Certain children develop all the changes of puberty, up to and including sperm production or ovulation, at an early age, either as the result of a brain lesion or as an isolated developmental, sometimes genetic, defect. The youngest mother on record was such a child; she gave birth to a full-term healthy infant by cesarean section at the age of five years and eight months. The existence of precocious puberty and the results of accidental ingestion by small children of male or female sex hormones indicate that breasts, uterus, and penis will respond to hormonal stimulation long before puberty. Evidently an increased end-organ sensitivity plays at most a minor part in puberal events.

The role of the hypothalamus in puberty

The signal to start the sequence of events is given by the brain, not the pituitary. Just as the brain holds the information on sex, so it holds information on maturity. The pituitary gland of a newborn rat successfully grafted in place of an adult pituitary begins at once to function in an adult fashion and does not have to wait until its normal age of maturation has been reached. It is the hypothalamus in the brain, not the pituitary, that must mature before puberty begins. Small amounts of sex hormones circulate from the time of birth, and these appear to inhibit the prepuberal hypothalamus from producing gonadotropin releasers. At puberty the hypothalamic cells become less sensitive to sex hormones. The small amount of sex hormones circulating then fails to inhibit the hypothalamus; gonadotropins are released, and these stimulate the production of testosterone by the testis or estrogen by the ovary. The level of the sex hormone rises until the same feedback circuit is re-established but now at a higher level of gonadotropins and sex hormones. The sex hormones are now high enough to stimulate the growth of secondary sex characters and to support mating behaviour.

Numerous factors may retard maturation or prevent normal growth, including hormonal disorders, metabolic defects, hereditary conditions, and inadequate nutrition. For information about specific growth disorders, see ENDOCRINE SYSTEMS: *Diseases and disorders of the human endocrine system*; METABOLISM; NUTRITION: *Nutritional diseases and disorder.* (J.M.T./Ed.)

AGING AND SENESCENCE

Life-span

Definition of life-span

It is a commonplace that organisms die, some after only a brief existence, like that of the mayfly whose adult life burns out in a day, and others like that of the gnarled bristlecone pines that have lived thousands of years. The limits of the life-span of each species appear to be determined ultimately by heredity. Locked within the code of the genetic material are instructions that specify the age beyond which a species cannot live given even the most favourable conditions. And many environmental factors act to diminish that upper age limit.

MEASUREMENT OF LIFE-SPAN

The maximum life-span is a theoretical number whose exact value cannot be determined from existing knowledge about an organism; it is often given as a rough estimate based on the longest lived organism of its species known to date. A more meaningful measure is the average life-span; this is a statistical concept that is derived by the analysis of mortality data for populations of each species. A related term is the expectation of life, a hypothetical number computed for humans from mortality tables drawn up by insurance companies. Life expectancy represents the average number of years that a group of persons, all born at the same time, might be expected to live, and it is based on the changing death rate over many past years.

The concept of life-span implies that there is an individual whose existence has a definite beginning and end. What constitutes the individual in most cases presents no problem: among organisms that reproduce sexually the in-

dividual is a certain amount of living substance capable of maintaining itself alive and endowed with hereditary features that are in some measure unique. In some organisms, however, extensive and apparently indefinite growth takes place and reproduction may occur by division of a single parent organism, as in many protists, including bacteria, algae, and protozoans. If these divisions are incomplete, a colony results; if the parts separate, genetically identical organisms are formed. In order to consider life-span in such organisms, the individual must be defined arbitrarily since the organisms are continually dividing. In a strict sense, the life-spans in such instances are not comparable to those forms that are sexually produced.

The beginning of an organism can be defined by the formation of the fertilized egg in sexual forms; or by the physical separation of the new organism in asexual forms (many invertebrate animals and many plants). In animals generally, birth is considered to be the beginning of the life-span. The timing of birth, however, is so different in various animals that it is only a poor criterion. In many marine invertebrates the hatchling larva consists of relatively few cells, not nearly so far along toward adulthood as a newborn mammal. For even among mammals, variations are considerable. A kangaroo at birth is about an inch long and must develop further in the pouch, hardly comparable to a newborn deer, who within minutes is walking about. If life-spans of different kinds of organisms are to be compared, it is essential that these variations be accounted for. The end of an organism's existence results when irreversible changes have occurred to such an extent that the individual no longer actively retains its organization. There is thus a brief period during which it is impossible to say whether the organism is still alive, but this time is so short relative to the total length of life that it creates no great problem in determining life-span.

Potential immortality

Some organisms seem to be potentially immortal. Unless an accident puts an end to life, they appear to be fully capable of surviving indefinitely. This faculty has been attributed to certain fishes and reptiles, which appear to be capable of unlimited growth. Without examining the various causes of death in detail (see DEATH) a distinction can be made between death as a result of internal changes (*i.e.*, aging) and death as a result of some purely external factor, such as an accident. It is notable that the absence of aging processes is correlated with the absence of individuality. In other words, organisms in which the individual is difficult to define, as in colonial forms, appear not to age. (P.W.F.)

Table 3: Maximum Ages for Some Seed Plants

plant	maximum age in years		locale of verified specimen
	estimated	verified	
Conifers			
Common juniper (*Juniperus communis*)	2,000	544	Kola Peninsula, northeastern Russia
Norway spruce (*Picea abies*)	1,200	350–400	Eichstätt, Bavaria
European larch (*Larix decidua*)	700	417	Riffel Alp, Switz.
Scotch pine (*Pinus sylvestris*)		584	
Swiss stone pine (*Pinus cembra*)	1,200	750	Riffel Alp, Switz.
White pine (*Pinus strobus*)		400–450	
Bristlecone pine (*Pinus aristata*)		4,900	Wheeler Peak, Humboldt National Forest, Nevada
Sierra redwood (*Sequoiadendron giganteum*)	4,000	2,200–2,300	Northern California
Flowering plants			
Monocots			
Dragon tree (*Dracaena draco*)	200*		Tenerife, one of the Canary Islands
Solomon's-seal (*Polygonatum*)	17†		
Dicots			
Dwarf birch (*Betula nana*)		80	Eastern Greenland
European beech (*Fagus sylvatica*)	900	250	Montigny, Normandy, France
English oak (*Quercus robur*)	2,000	1,500	Hasbruch Forest, Lower Saxony
Bo tree (*Ficus religiosa*)	2,000–3,000‡		Buddh Gaya, India; Anuradhapura, Ceylon
Linden (*Tilia*)		815	Lithuania
English ivy (*Hedera helix*)	440		Ginac, near Montpellier, France

*Exaggerated estimates for this historic specimen reach 6,000 years. †Scars on root-stock counted. ‡According to Buddhist and Roman history.

PLANTS

Plants grow old as surely as do animals; however, a generally accepted definition of age in plants has not yet been realized. If the age of an individual plant is that time interval between the reproductive process that gave rise to the individual and the death of the individual, the age attained may be given readily for some kinds of plants but not for others. Table 3 lists maximum ages, both estimated and verified, for some seed plants.

Problem of defining age. An English oak that has 1,000 annual rings in the trunk is 1,000 years old. But age is less certain in the case of an arctic lupine that germinated from a seed that, containing the embryo, had been lying in a lemming's burrow in the arctic permafrost for 10,000 years.

The mushroom caps that appear overnight last for only a few days, but the network of fungus filaments in the soil (the mycelia) may be as old as 400 years. Because of important differences in structure, the life-span of higher plants cannot be compared with that of higher animals. Normally, embryonic cells (that is, cells capable of changing in form or becoming specialized) cease to exist very early in the life of an animal. In plants, however, embryonic tissue—the plant meristems—may contribute to growth and tissue formation for a much longer time, in some cases throughout the life of the plant. Thus the oldest known trees, bristlecone pines of California and Nevada, have one meristem (the cambium) that has been adding cells to the diameter of these trees for, in many cases, more than 4,000 years and another meristem (the apical) that has been adding cells to the length of these trees for the same period. These meristematic tissues are as old as the plant itself; they were formed in the embryo. The wood, bark, leaves and cones, however, live for only a few years. The wood of the trunk and roots, although dead, remains a part of the tree indefinitely, but the bark, leaves, and cones are continually in the process of dying and sloughing off.

Ageless embryonic tissue

Among the lower plants only a few mosses possess structures that enable an estimate of their age to be made. The haircap moss (*Polytrichum*) grows through its own stem tip each year, leaving a ring of scales that marks the annual growth. Three to five years' growth in this moss is common, but life-spans of 10 years have been recorded. The lower portions of such a moss are dead, though intact. Peat moss (*Sphagnum*) forms extensive growths that fill acid bogs with a peaty turf consisting of the dead lower portions of mosses whose living tops continue growing. Mosses that become encrusted with lime (calcium carbonate) and form "tufa" beds several metres thick also have living tips and dead lower portions. On the basis of their observed annual growth, some tufa mosses are estimated to have been growing for as long as 2,800 years.

No reliable method for determining the age of ferns exists, but on the basis of size attained and growth rate, some tree ferns are thought to be several decades old. Some club mosses, or lycopsids, have a "storied" growth pattern similar to that of the haircap moss. Under favourable conditions some specimens live five to seven years.

The woody seed plants, such as conifers and broadleaf trees, are the most amenable to determination of age. In temperate regions, where each year's growth is brought to an end by cold or dryness, every growth period is limited by an annual ring—a new layer of wood added to the diameter of the tree. These rings may be counted on the cut ends of a tree that has been felled or, using a special instrument, a cylinder of wood can be cut out and the growth rings counted and studied. In the far north growth rings are so close together that they are difficult to count.

In the moist tropics growth is more or less continuous, so that clearly defined rings are difficult to find.

Errors in estimating age

Often the age of a tree is estimated on the basis of its diameter, especially when the average annual increase in diameter is known. The source of greatest error in this method is the not infrequent fusing of the trunks of more than one tree, as, for example, occurred in a Montezuma cypress in Santa María del Tule, a little Mexican village near Oaxaca. This tree, described by the Spanish explorer Hernan Cortés in the early 1500s, was earlier estimated on the basis of its great thickness to be 6,000 years old; later studies, however, proved it to be three trees grown together. Estimates of the age of some English yews have been as high as 3,000 years, but these figures, too, have turned out to be based on the fusion of close-growing trunks, none of which is more than 250 years old. Increment borings of bristlecone pines have shown specimens in the western United States to be 4,600 years old.

Growing season of seed plants. *Annuals.* Plants, usually herbaceous, that live for only one growing season and produce flowers and seeds in that time are called annuals. They may be represented by such plants as corn and marigolds, which spend a period of a few weeks to a few months rapidly accumulating food materials. As a result of hormonal changes—brought about in many plants by changes in environmental factors such as day length and temperature—leaf-producing tissues change abruptly to flower-producing ones. The formation of flowers, fruits, and seeds rapidly depletes food reserves and the vegetative portion of the plant usually dies. Although the exhaustion of food reserves often accompanies death of the plant, it is not necessarily the cause of death.

Biennials. These plants, too, are usually herbaceous. They live for two growing seasons. During the first season, food is accumulated, usually in a thickened root (beets, carrots); flowering occurs in the second season. As in annuals, flowering exhausts the food reserves, and the plants die after the seeds mature.

Perennials. These plants have a life-span of several to many years. Some are herbaceous (iris, delphinium), others are shrubs or trees. The perennials differ from the above-mentioned groups in that the storage structures are either permanent or are renewed each year. Perennials require from one to many years growth before flowering. The preflowering (juvenile) period is usually shorter in trees and shrubs with shorter life-spans than in those with longer life-spans. The long-lived beech tree (*Fagus sylvatica*), for example, passes 30–40 years in the juvenile stage, during which time there is rapid growth but no flowering.

Some plants—cotton and tomatoes, for example—are perennials in their native tropical regions but are capable of blooming and producing fruits, seeds, or other useful parts in their first year. Such plants are often grown as annuals in the temperate zones.

Longevity of seeds. Although there is great variety in the longevity of seeds, the dormant embryo plant contained within the seed will lose its viability (ability to grow) if germination fails to occur within a certain time. Reports of the sprouting of wheat taken from Egyptian tombs are unfounded, but some seeds do retain their viability a long time. Indian lotus seeds (actually fruits) have the longest known retention of viability. On the other hand, seeds of some willows lose their ability to germinate within a week after they have reached maturity.

The oldest living seeds

The loss of viability of seeds in storage, although hastened or retarded by environmental factors, is the result of changes that take place within the seed itself. The changes that have been investigated are: exhaustion of food supply; gradual denaturing or loss of vital structure by protoplasmic proteins; breakdown of enzymes; accumulation of toxins resulting from the metabolism of the seed. Some self-produced toxins may cause mutations that hamper seed germination. Since seeds of different species vary greatly in structure, physiology, and life history, no single set of age factors can apply to all seeds. (L.K.)

ANIMALS

Much of what is known of the length of life of animals other than man derives from observations of domesticated species in laboratories and zoos. One has only to consider how few animals reveal their age to appreciate the difficulties involved in answering the apparently simple question of how long they live in nature. In many fishes, a few kinds of clams, and an occasional species of other groups, growth is seasonal, so that annual zones of growth, much like tree rings, are produced in some part of the organism. Among game species, methods of determining relative age by indicators such as the amount of tooth wear or changes in bone structure have yielded valuable information. Bird bands and other identifying marks also make age estimation possible. But one of the consequences of the fact that animals move is that very little is known about the life-span of most species as they exist in nature.

Maximum and average longevity. The extreme claims of longevity that are occasionally made for one species or another have consistently been proven false when subjected to critical scrutiny. Although the maximum life-span that has been observed for a particular species cannot be considered absolute, since a limited number of individuals at best has been studied, this datum probably provides a fair approximation of the greatest age attainable for this kind of animal under favourable conditions. Animals in captivity, which provide most of the records of extreme age, are exposed to far fewer hazards than those in the wild. In the accompanying table of maximum longevity, particular species have been so selected as to encompass the known range of longevity of other members of the taxonomic group to which they belong.

Table 4: Maximum Longevity of Animals in Captivity*

animal	life-span in years	animal	life-span in years
Mammals		**Amphibians**	
Bat (*Eptesicus fuscus*)	2	European black salamander (*Salamandra atra*)	3
Grizzly bear (*Ursus horribilis*)	31	Spotted salamander (*Ambystoma maculatum*)	25
Cat (*Felis catus*)	21	Frog (*Rana* species)	5–15
Chimpanzee (*Pan troglodytes*)	37	**Fishes**	
Dog (*Canis familiaris*)	34	Eel (*Anguilla rostrata*)	6
Elephant, Indian (*Elephas maximus*)	57	Goldfish (*Carassius auratus*)	25
Goat (*Capra hircus*)	18	Sturgeon (*Acipenser transmontanus*)	50
Golden hamster (*Mesocricetus auratus*)	1.8	**Insects**	
Horse (*Equus caballus*)	62	Ant (*Lasius* species)	15
Lion (*Panthera leo*)	29	Buprestid beetle (*Buprestis splendens*)	30
Mouse (*Mus musculus*)	3	Fruit fly (*Drosophila melanogaster*)	0.1
Ox (*Bos taurus*)	30	**Arachnids**	
Squirrel, gray (*Sciurus carolinensis*)	15	Bird spider (*Avicularis avicularis*)	15
Wild boar (*Sus scrofa*)	27	Rocky Mountain wood tick (*Dermacentor andersoni*)	3–4
Birds		**Crustaceans**	
Blue jay (*Cyanocitta cristata*)	4	Crayfish (*Astacus fluviatilis*)	30
Canary (*Serinus canaria*)	24	Water flea (*Daphnia magna*)	0.2
Macaw (*Ara macao*)	64	**Mollusks**	
Nightingale (*Luscinia luscinia*)	3.8	Clams, various species	1–10
Pigeon (*Columba livia domestica*)	35	Snails, various species	1–30
Titmouse (*Parus major*)	9	**Annelid worms**	
Reptiles		Earthworm (*Lumbricus terrestris*)	10
Alligator (*Alligator mississippiensis*)	56	Medicinal leech (*Hirudo medicinalis*)	27
Garter snake (*Thamnophis sirtalis*)	6	**Rotifers**	
Box turtle (*Terrapene carolina*)	123	Various species	0.03–0.1
Giant tortoise (*Testudo elephantopus*)	177		
Water turtle (*Pseudemys scripta*)	7		

*Condensed and adapted from W.S. Spector (ed.), *Handbook of Biological Data*, 1956.

Environmental influences. Life-span usually is measured in units of time. Although this may seem eminently logical, certain difficulties may arise. In cold-blooded animals in general, the rate of metabolism that determines the various life processes varies with the temperatures to which they are exposed. If aging depends on the expenditure of a fixed amount of vital energy, an idea first proposed in 1908, life-span will vary tremendously depending on

Temperature effects

temperature or other external variables that influence life-span. There is considerable evidence attesting at least to the partial cogency of this argument. So long as a certain range is not exceeded, cold-blooded invertebrates do live longer at low than at high temperatures. Rats in the laboratory live longest on a somewhat restricted diet that does not permit maximum metabolic rate. Of perhaps even greater significance is the fact that many animals undergo dormant periods. Many small mammals hibernate; a number of arthropods have life cycles that include periods during which development is arrested. Under both conditions the metabolic rate becomes very low. It is questionable whether such periods should be included in computing the life-span of a particular organism. Comparisons between species, some of which have such inactive periods while others do not, are dangerous. It is possible that life-span could be measured more adequately by total metabolism; however, the data that are necessary for this purpose are almost entirely lacking.

Length of life is controlled by a multitude of factors, which collectively may be termed environment, operating on a genetic system that determines how the individual will respond. It is impossible to list all the environmental factors that may lead to death. For analytical purposes it is, however, useful to make certain formal separations. Every animal is exposed to (1) a pattern of numerous events, each with a certain probability of killing the individual at any moment and, in the aggregate, causing a total probability of death or survival; (2) climatic and other changes in the habitat, modifying the frequency with which the various potentially fatal events occur; and (3) progressive systemic change, inasmuch as growth, reproduction, development, and senescence are characteristics intrinsic in the organism and capable of modifying the effects of various environmental factors.

Patterns of survival. Consider a group of similar animals of the same age. Although no two individuals can have precisely the same environment, let it be assumed that the environment of the group remains effectively constant. If the animals undergo no progressive physiological changes, the factors causing death will produce a death rate that will remain constant in time. Under these conditions, it will take the same amount of time for the population to become reduced to one-half its former number, no matter how many animals remain at the beginning of the period considered. The animals therefore survive according to the pattern of an accident curve. This is the sense in which many of the lower animals are immortal. Although they die, they do not age; how long they have already lived has no influence on their further life expectation.

Another group of animals may consist of individuals that differ markedly in their responses to the constant environment. They may be genetically different, or their previous development may have caused variations to arise. Those individuals that are most poorly suited to the new environment will die, leaving survivors that are better adapted. The same result can also be achieved in other ways. If the environment varies geographically, those individuals that happen to find areas in which existence can be maintained will survive, while the remainder will die. Or, as a result of their own properties, animals in a constant environment may acclimate in a variety of ways, thus adjusting to the existing conditions. The pattern of survival that results in each of these cases is one in which the death rate declines with time, as illustrated by the selection–acclimation curve.

In the absence of death from other causes, all members of a population may exist in their environment until the onset of senescence, which will cause a decline in the ability of individuals to survive. In a sense they can be considered to wear out as does a machine. Their survival is best described by individual differences among members of the population that determine the curvature of the survival line (wearing-out curve). The more the population varies, the less abrupt is the transition from total survival to total death.

Under the actual conditions of existence of animals the three types of survival (accident pattern, selection–acclimation pattern, wearing-out pattern) above all enter as components of the realized survival pattern. Thus in animals that are carefully maintained in the laboratory, survival is approximately that of the wearing-out pattern. Environmental accidents can be kept to a minimum under these conditions, and survival is almost complete during the major part of the life-span. In all known cases, however, the early stages of the life-span are characterized by a noticeable contribution of the selection–acclimation pattern. This must be interpreted as a result of developmental changes that accompany the early life of the individuals and of selective processes that operate on those organisms whose genetic constitutions are ill fitted for that environment.

In some of the larger mammals in nature, the existing evidence points to a similar survival pattern. In a variety of other animals, however, and including fishes and invertebrates, mortality in the young stages is so high that the selection–acclimation curve predominates. One estimate places the mortality of the Atlantic mackerel during its first 90 days of life as high as 99.9996 percent. Since some mackerel do live for several years, a mortality rate that decreases with age is indicated. Similar considerations probably apply to all those animals that have larval stages that serve as dispersal mechanisms.

Highest mortality in aquatic animals

When the postjuvenile portion of the life-span is considered by itself, a number of animals for which such information has been gathered—including primarily fishes and birds—have survivorship curves that are dominated by the accident pattern. In these species in nature, death from old age apparently is rare. Their chance of surviving to an advanced age is so small that it may be statistically negligible. In modern times, human predation is a large factor in the mortality of these species in many cases. Since deaths from fishing and hunting are largely independent of age, once an animal has reached a certain minimum size, such a factor only makes the survival curve steeper but does not change its shape. One consequence of such increased mortality is that fewer old and large individuals are noticed in a population.

More complex survival patterns, such as the hypothetical one illustrated, undoubtedly exist. They should be looked for in those species in which extensive reorganization of the animal is part of the normal life cycle. In effect, these animals change their environment radically, in some cases several times during a lifetime. The frog offers a familiar example. During its period of early development and until shortly after hatching, the animal is subject to major internal, and some external, change. As a tadpole it is adjusted to an aquatic, herbivorous life. The metamorphosis to the terrestrial, carnivorous adult form is accompanied by varied physiological stresses that must be expected to produce a temporary increase in mortality rate. In some insects the eggs, larvae, pupae, and adults are exposed to and respond to quite different environments, and a survivorship pattern even more complex than that described by the composite curve may exist.

Effects of metamorphosis

The same species will exhibit changed survival in different environments. In captivity an animal population may approach the wearing-out pattern; in its natural habitat survivorship may vary with age in a quite different way. Although one can assign a maximum potential life-span to an individual—while realizing that this maximum may not be attained—it is impossible to specify the survivorship pattern unless the environment is also specified. This is another way of saying that life-span is the joint property of the animal and the environment in which it lives.

HUMAN LIFE-SPAN

The exact duration of human life is unknown, although there is presumably a maximum life-span for the human race established in the genetic material. At first thought, this statement seems irrational. Surely no human being can live 1,000 years. Even though all may agree that the likelihood of an individual living 1,000 years is infinitesimal, there is no scientific proof that this statement is or is not true. The indeterminacy of the maximum limit of human life is made more comprehensible if one chooses a number that may appear to be a more reasonable limit.

Since there is no verified instance of a person having

lived 150 years, this number may, for purposes of illustration, be arbitrarily accepted as the maximum limit of the span of human life. But if the possibility is admitted that an individual may live exactly 150 years, there is no valid reason for rejecting the possibility that some other individual may live 150 years and one minute. And if 150 years and one minute is accepted, why not 150 years and two minutes, and so on? Thus, based on existing knowledge of longevity, a precise figure for the span of human life cannot be given.

Studies on longevity. Much information concerning the inheritance of longevity has come from the study of genealogical records of nobility and landed gentry. The early genealogical studies were criticized on the grounds that the downward trend in the death rate (attributable generally to scientific advancements) introduced a spurious correlation in statistics derived from records extending over long periods of time. It was argued that in some instances records were included of persons who, at the time of the study, had not had the opportunity of living out their possible life-span. The general finding of such investigations was that the expectation of life of sons of long-lived parents (*i.e.*, those living to age 70 years or older) was greater than that of sons of shorter-lived parents (*i.e.*, those having attained less than age 50 at the time of death).

An American biostatistician attempted to avoid the defects of genealogical studies by collecting records of the family histories of 365 nonagenarians (90-year-old persons) and of a comparison group of 143 individuals of varying ages, selected because all of their six immediate ancestors were dead. The study introduced the concept of "total immediate ancestral longevity," or TIAL—the sum of the ages at death of the two parents and the four grandparents of a given person—as a measure of longevity. This number is unlikely to be greater than 600 or less than 90. The average TIAL of the nonagenarians and centenarians definitely exceeded that of the comparison group. This held true not only for the six immediate ancestors as a group but also for each category—father, mother, paternal and maternal grandparents. In the same study, investigators also computed the expectation of life for sons of fathers as classified in three groups by age at death: (1) under age 50, (2) from age 50 to age 79, and (3) age 80 or over. The expectation of life for the three groups at birth was 47.0, 50.5, and 57.2 years, respectively. The same relative ranking continued through the lifetime of the sons, their expectation of life at age 40 being 27.3, 28.9, and 32.0 years, respectively.

The TIAL concept

While certain doubts have been raised about the validity of these as well as earlier studies, taken at their face value, these data show clearly that long-lived persons had parents and grandparents who lived longer than the parents and grandparents of shorter lived persons.

Since longevity is important in life insurance underwriting, several studies have been made of the relationship between heredity and the life-span by an analysis of life insurance records. Such analyses showed that policyholders both of whose parents were living when the policy was written live longer than those whose parents were dead when the policy was written. These results are in conformity with those obtained from genealogical records and family histories.

Influence of heredity

Each of the various types of studies of the inheritance of longevity—genealogical records, life insurance records, and family histories of the general population—has limitations that restrict the applicability of the findings. The principal studies indicate, nevertheless, that the children of long-lived parents are more likely to be long-lived than are the children of short-lived parents. Conversely, the immediate ancestors—parents and grandparents—of long-lived persons on the average are older at death than are the immediate ancestors of persons who die at a relatively young age. These studies support the conclusion, mentioned earlier, that longevity is determined in part by heredity.

Actual versus possible life-span. It should be observed that this conclusion relates to the inheritance of longevity—the observed expression of the span of life—and not to the span of life itself. The actual length of life itself is shorter than the possible life-span, since the former reflects the effect of unfavourable environmental factors. In the absence of any biological data from which the maximum limit of the span of life can be determined precisely, an estimate of the limit must be obtained from observation of the actual length of life of persons who already have died. But such observations cannot establish a fixed limit for the span of life.

The estimation of the length of the span of life from observed data is a form of sampling from a large but incomplete population. The tabulation of the ages at death of a large number of persons from a large general population of the United States will give an asymmetrical frequency distribution with two modes, or peaks, of highest frequency: the first at age less than one year and the second between ages 75 and 80 years. The frequency distribution is bounded by age zero at the lower limit but there is no boundary at the upper limit. The number of deaths of persons whose length of life is near the upper limit of this frequency distribution (*e.g.*, 100 years or more) varies from year to year. The age of the oldest person dying also varies from year to year.

The number of deaths of centenarians (100-year-old persons) depends in part upon the number of deaths counted. Ages at death are frequently unverified, so that the true numbers of centenarians almost certainly deviate from those given in official vital statistics. Moreover, only a very small proportion of the deaths that have occurred throughout the history of the human race have been registered. The potential number of future deaths greatly exceeds the number that already has occurred. Statistical theory supports the expectation that as the total number of deaths continues to increase, the death of a person whose length of life will be longer than that of any person previously known will be recorded.

Observation of the length of life of persons who have died can show that it is possible for a human being to live to the oldest age recorded as of any specified date and can provide an estimate of the relative frequency or probability of that event. But such observations do not provide a logical basis for fixing any age as the maximum possible limit of the life-span.

The continuation of the worldwide decline in the death rate will naturally result in an increase in the number of persons who live until age 100 years or more. Since the number of persons who may live to an advanced age, such as 110 or 115 years, is directly related to the number of persons who live to age 100, an increase in the latter number will increase the probability that the death of an individual attaining some greater age (*e.g.*, 115 years) will be recorded at some future date.

Many instances of persons alleged to have died at an age considerably greater than 100 years have been recorded. Statements concerning the age at death of biblical characters such as Methuselah can be dismissed, since scientific verification is impossible. Three of the most frequently cited cases of more recent times are: Thomas Parr, who died in November 1635 at the alleged age of 152 years; Henry Jenkins, who died in December 1670 at the alleged age of 169 years; and Catherine, countess of Desmond, who died in 1604 at the alleged age of 140 years. William Harvey, a famous English physician, performed an autopsy on Thomas Parr and the account of the autopsy was cited for many years as evidence that Harvey—in his paper—had confirmed Parr's age. Quite apart from the fact that it is impossible accurately to determine the age of a person by an autopsy, Harvey made no attempt to verify Parr's age but merely referred to the current estimates. Subsequent investigations have revealed that no proof exists of the age at death of any of these three individuals and that their reported ages were based solely upon hearsay.

Famous age records

An example with more definite documentation is that of Christian Jacobsen Drakenberg, stated to have been born on November 18, 1626, and to have died on October 9, 1772, aged 145 years and 325 days. Although the authenticity of his age was attested to by many persons, including two celebrated Scandinavian actuaries, later investigations cast doubt upon the record. It is difficult to accept the statements concerning Drakenberg's age at death, since

this age is more than 30 years greater than the next oldest verified age at death—a difference that in itself casts doubt on its authenticity.

Of eight individuals for whom records substantiate the fact that each had lived more than 108 years, seven were females. Six of the eight were more than 110 years old at death. The oldest was Pierre Joubert, who was born July 15, 1701, and died November 16, 1814, aged 113 years and 124 days. Discounting the Drakenberg record, this is the oldest age at death that has been generally accepted as authentic.

It may be concluded that the span of human life is at least 114 years, but that this is not the maximum upper limit. This does not mean the span of life of each individual now living or to be born in the future is at least 114 years. The span of life, since it is determined by heredity, varies from one individual to another as do other genetically determined traits.

Factors influencing life-span of man

A significant proportion of human embryos and fetuses die before birth. Other infants at birth have defects that limit their span of life to a few years. Some malformations (*e.g.*, certain cardiovascular defects) are developmental rather than genetic in the strict sense of the word and can be corrected so that the length of life of such persons is extended.

In the past the length of life of most individuals has been considerably shorter than their possible span of life because of unhealthful environmental factors. As these factors are increasingly brought under control or eliminated, the actual length of life will approach more closely the span of life. At the end of the 18th century the expectation of life at birth in North America and northwestern Europe was about 35 or 40 years. By 1970 it exceeded 70 years, and at some future date the death of a person at an authenticated age of more than 114 years can be expected.

There is no evidence that the span of human life has increased since the beginning of recorded history. Neither is there any evidence that the death rate of centenarians has decreased. The expected increase in the number of centenarians results from a decrease in the death rate at ages under 100 years and not from any demonstrable increase in the maximum length of the span of life. The remarkable increase in the average length of life during the past 2,000 years—from 20–25 years to 70 years under favourable conditions—has increased the likelihood that a person may live to the maximum limit of his span of life. (P.W.F.)

Aging: general considerations

Aging consists of the progressive changes that take place in a cell, an organ, or the total organism with the passage of time. It is a process that goes on over the entire adult life-span of any living thing. Gerontology, the study of the aging process, is devoted to the understanding and control of all factors contributing to the finitude of individual life. It is not concerned exclusively with debility, which looms so large in human experience, but deals with a much wider range of phenomena. Every species has a life history in which the individual life-span has an appropriate relationship to the reproductive life-span and to the mechanism of reproduction and the course of development. How these relationships evolved is as germane to gerontology as it is to evolutionary biology. It is also important to distinguish between the purely physicochemical processes of aging and the accidental organismic processes of disease and injury that lead to death.

Gerontology, therefore, can be defined as the science of the finitude of life as expressed in the three aspects of longevity, aging, and death, examined in both evolutionary and individual (ontogenetic) perspective. Longevity is the span of life of an organism. Aging is the sequential or progressive change in an organism that leads to an increased risk of debility, disease, and death; senescence consists of these manifestations of the aging process.

Actuarial functions

The viability (survival ability) of a population is characterized in two actuarial functions: the survivorship curve (A in Figure 39) and the age-specific death rate, or Gompertz function (B in Figure 39). The relation of such factors as aging characteristics, constitutional vigour, physical factors, diet, and exposure to disease-causing organisms to the actuarial functions is complex; there is, nevertheless, no substitute for them as measures of the aging process and of the effect of environmental or genetic modifiers.

The age-specific mortality rate is the most informative actuarial function for investigations of the aging process. It was first pointed out by an English actuary, Benjamin Gompertz, in 1825 that the mortality rate increases in geometric progression—*i.e.*, by a constant ratio in successive equal age intervals. Hence, a straight line, known as the Gompertz function, results when death rates are plotted on a logarithmic (ratio) scale. The prevalence of many diseases and disabilities rises in the same geometrical manner as does the mortality rate, important exceptions being some infectious diseases and diseases arising from disturbances of the immunological system. Although the life tables of most species are remarkably similar in form, even closely related species can differ markedly in the relative incidence of the major causes of death. (G.A.Sa.)

BIOLOGICAL THEORIES OF AGING

Aging has many facets. Hence there are a number of theories, each of which many explain one or more aspects of aging; there is, however, no single theory that explains all of the phenomena of aging.

Genetic theories. One theory of aging assumes that the life-span of a cell or organism is genetically determined—that the genes of an animal contain a "program" that determines its life-span just as eye colour is determined genetically. Although long life is recognized often as a familial characteristic, and short-lived strains of fruit flies, rats, and mice can be produced by selective breeding, other factors clearly can significantly alter the basic genetic program of aging.

Another genetic theory of aging assumes that cell death is the result of "errors" introduced in the formation of key proteins, such as enzymes. Slight differences induced in the transmission of information from the deoxyribonucleic acid (DNA) molecules of the chromosomes through ribonucleic (RNA) molecules (the "messenger" substance) to the proper assembly of the large and complex enzyme molecules could result in a molecule of the enzyme that would not "work" properly. These so-called error theories have not yet been firmly established, but studies are in progress.

Cell errors and mutations

As cells grow and divide, a small proportion of them undergo mutation; that is, they become "different" with a change in their chromosome structure that is then reproduced when they again divide. The "somatic mutation" theory of aging assumes that aging is due to the gradual accumulation of mutated cells that do not perform normally.

Non-genetic theories. Other theories of aging focus attention on factors that can influence the expression of a genetically determined "program." One of these is the "wear-and-tear" theory, which assumes that animals and cells, like machines, simply wear out. Animals, however, unlike machines, have some ability to repair themselves, so that this theory does not fit the facts of a biological system. A corollary to the wear-and-tear theory is the presumption that waste products accumulate within cells and interfere with function. The accumulation of highly insoluble particles, known as "age pigments," has been observed in muscle cells in the heart and nerve cells of both human beings and other animals.

With increasing age, tendons, skin, and even blood vessels lose elasticity. This is due to the formation of cross-links between or within the molecules of collagen (a fibrous protein) that give elasticity to these tissues. The "cross-linking" theory of aging assumes that similar cross-links form in other biologically important molecules, such as enzymes. These cross-links could alter the structure and shape of the enzyme molecules so that they are unable to carry out their functions in the cell.

Auto-immunity control theories

Another theory of aging assumes that immune reactions, normally directed against disease-producing organisms as well as foreign proteins or tissue, begin to attack cells of the individual's own body. In other words, the system that produces antibodies loses its ability to distinguish between

"self" and foreign proteins. This "autoimmune" theory of aging is based on clinical rather than on experimental evidence.

These theories all attempt to explain aging in terms of cellular and molecular changes. Actually, age changes are much more marked in the overall performance of an individual than in cellular processes that can be measured. The age decrement in the ability to perform muscular work is much greater than any changes that can be detected in the enzyme activities of the muscles that perform the work. It is possible that aging in an individual is actually due to a breakdown in the control mechanisms that are required in a complex performance. (N.W.S./Ed.)

NATURAL HISTORY OF AGING

Reproduction and aging. Reproduction is an all-important function of an organism's life history, and all other vital processes, including senescence and death, are shaped to serve it. The distinction between semelparous and iteroparous modes of reproduction is important for an understanding of biological aging. Semelparous organisms reproduce by a single reproductive act. Annual and biennial plants are semelparous, as are many insects and a few vertebrates, notably salmon and eels. Iteroparous organisms, on the other hand, reproduce recurrently over a reproductive span that usually covers a major part of the total life-span.

In semelparous forms, reproduction takes place near the end of the life-span, after which there ensues a rapid senescence that quickly leads to the death of the organism. In plants the senescent phase is usually an integral part of the reproductive process and essential for its completion. The dispersal of seeds, for example, is accomplished by processes—including ripening and fall (abscission) of fruits and drying of seed pods—that are inseparable from the overall senescence process. Moreover, the onset of plant senescence is invariably initiated by the changing levels of hormones, which are under systemic or environmental control. If, for example, the hormone auxin is prevented, by experimental means, from influencing the plant, the plant lives longer than normal and undergoes an atypical prolonged pattern of senescent change.

The significance of aging patterns in insects

Useful inferences can be drawn from the study of the aging processes of insects that display two distinct kinds of adaptive coloration: the procryptic, in which the patterns and colours afford the insect concealment in its native habitat; and the aposematic, in which the vivid markings serve as a warning that the insect is poisonous or bad tasting. The two adaptation patterns have different optimal species survival strategies: the procryptics die out as quickly as possible after completing reproduction, thus reducing the opportunity for predators to learn how to detect them; the aposematics have longer post-reproductive survival, thus increasing their opportunity to condition predators. Both adaptations are found in the family of saturniid moths, and it has been shown that the duration of their post-reproductive survival is governed by an enzyme system that controls the fraction of time spent in flight: procryptics fly more, exhaust themselves, and die quickly; aposematics fly less, conserve their energies, and live longer.

These examples indicate that in semelparous forms, in which full vigour and function are required until virtually the end of life, senescence has an onset closely coupled with the completion of the reproductive process and is governed by relatively simple enzymatic mechanisms that can be modified by natural selection. Such specific, genetically controlled senescence processes are instances of programmed life termination.

The iteroparous forms include most vertebrates, most of the longer-lived insects, crustaceans and spiders, cephalopod and gastropod mollusks, and perennial plants. In contrast to semelparous forms, iteroparous organisms need not survive to the end of their reproductive phase in order to reproduce successfully, and the average fraction of the reproductive span survived varies widely between groups: small rodents and birds in the wild survive on the average only 10 percent to 20 percent of their potential reproductive lifetimes; whales, elephants, apes, and other large mammals in the wild, on the other hand, live through 50 percent or more of their reproductive spans, and a few survive beyond reproductive age. In iteroparous forms the onset of senescence is gradual, with no evidence of specific systemic or environmental initiating mechanisms; senescence manifests itself early as a decline in reproductive performance. In species that grow to a fixed body size, decline of reproductive capacity begins quite early and accelerates with increasing age. In large egg-laying reptiles, which attain sexual maturity while relatively small in size and continue to grow during a long reproductive span, the number of eggs laid per year increases with age and body size but eventually levels off and declines. The reproductive span in such cases is shorter than the life-span.

These comparisons illustrate the influence exerted by factors of population dynamics on the evolution of reproductive and bodily (somatic) senescence. The proportional contribution of an individual to the rate of increase of the iteroparous population obviously diminishes as the number of his living progeny increases. In addition, his reproductive capacity diminishes with age. These facts imply that there is an optimum number of litters per lifetime. Whether or not these influences of population dynamics lead to the evolution of adaptive senescence patterns has long been debated by gerontologists but has not yet been investigated definitively.

Maximum life-span as an indicator of longevity

Species differences in longevity and aging. That there are large differences in life-span between some species of animals has long been known, but only recently have the data become adequate for statistical analysis. Maximum life-span provides an estimate of the potential longevity of mammalian and avian species because of the sharp upper limit of the survival curves in life tables. Also, it is superior to the average life-span because the latter is influenced by environmental factors unrelated to aging (*e.g.*, human protection).

The taxonomic stratification of longevity can be seen among the mammals. Primates, generally, are the longest lived group, although some small prosimians and New World monkeys have relatively short life-spans. The murid (mouselike) rodents are short-lived; the sciurid (squirrel-like) rodents, however, can reach ages two to three times longer than the murids. Three traits have independent correlations with life-span: brain weight, body weight, and resting metabolic rate. The dependence of life-span on these traits can be expressed in the form of an equation: $L = 5.5E^{0.54}S^{-0.34}M^{-0.42}$. Mammalian life-span ($L$) in months relates to brain weight (E) and body weight (S) in grams and to metabolic rate (M) in calories per gram per hour. The positive exponent for E (0.54) indicates that longevity of mammals has a strong positive association with brain size, independent of body size or metabolic rate. The negative coefficient for metabolic rate implies that life-span decreases as the rate of living increases, if brain and body weight are held constant. The negative partial coefficient for body weight indicates that the tendency for large animals to be longer lived results not from body size but rather from the high positive correlation of body weight with brain weight and its negative correlation with metabolic rate. The same kind of relation of L to E, S, and M holds for birds, but there is a tendency for birds to be longer lived than mammals of comparable brain and body size despite their higher body temperatures and metabolic rates. The larger reptiles have life-spans exceeding those of mammals of comparable size, but their rates of metabolism are about ten times lower, so that their total lifetime energy expenditures are lower than those for mammals. The more highly cephalized animals (*i.e.*, those with higher brain weight), especially the primates, have greater lifetime energy outputs; the total lifetime energy output per gram of tissue is about 1,200,000 calories for man and 400,000 calories for domestic animals such as cats and dogs.

The above relations hold for the homeothermic mammals, those with nearly constant body temperature. The heterothermic mammals, which are able to enter daily torpor, or seasonal hibernation, thereby reduce their metabolic rates more than tenfold. The insectivorous bats of temperate latitudes are the most dramatic example;

although they have life-spans in excess of 20 years, almost 80 percent of that time is spent in deep torpor. As a result, their lifetime energy expenditures are no greater than are those of other small mammals.

The longevities of arthropod species extend from a few days to several decades. The extremely short-lived insects have a brief single reproductive phase; the longer lived spiders and crustaceans are iteroparous, with annual reproductive cycles.

The inheritance of longevity. The inheritance of longevity in animal populations such as fruit flies and mice is determined by comparing the life tables of numerous inbred populations and some of their hybrids. The longevity of sample populations has been measured for more than 40 inbred strains of mice. Two experiments concur in finding that about 30 percent of longevity variation in female mice is genetically determined, whereas the heritability in male mice is about 20 percent. These values are comparable to the heritabilities of some physiological performances, such as lifetime egg or milk production, in domestic animals.

Rate of aging

The slope of the Gompertz function line indicates the rate of actuarial aging. The differences in longevity between species are the result primarily of differences in the rate of aging and are therefore expressed in differences in slope of the Gompertz function. Three species that vary in longevity, such as the opossum (seven years), rabbit (12 years), and cat (20 years), would have Gompertz functions like the three lines in B (see Figure 39), and they would have survival curves like the corresponding curves in 39A.

By courtesy of Argonne National Laboratory, Argonne, Illinois

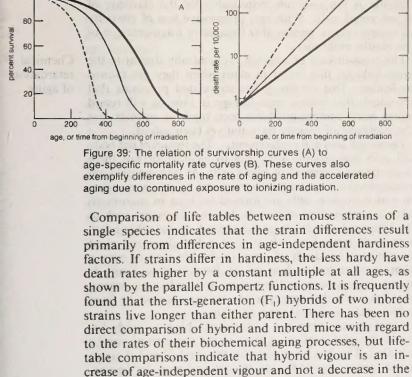


Figure 39: The relation of survivorship curves (A) to age-specific mortality rate curves (B). These curves also exemplify differences in the rate of aging and the accelerated aging due to continued exposure to ionizing radiation.

Comparison of life tables between mouse strains of a single species indicates that the strain differences result primarily from differences in age-independent hardiness factors. If strains differ in hardiness, the less hardy have death rates higher by a constant multiple at all ages, as shown by the parallel Gompertz functions. It is frequently found that the first-generation (F_1) hybrids of two inbred strains live longer than either parent. There has been no direct comparison of hybrid and inbred mice with regard to the rates of their biochemical aging processes, but life-table comparisons indicate that hybrid vigour is an increase of age-independent vigour and not a decrease in the rate of aging.

Recent research indicates that much of the variation in survival time between mouse strains is attributable to differences in inherited susceptibility to specific diseases. An important task of gerontology is to determine the extent of such genetic influences on aging.

The inheritance of longevity in humans is more difficult to investigate because length of life is influenced by socio-economic and other environmental factors that generate spurious correlations between close relatives. A number of studies have been published, most of them pointing to some degree of heritability with regard to length of life or susceptibility to major diseases, such as cancer and heart disease. Although there is disagreement about the degree of heritability of longevity in man, the evidence for genetic transmission of susceptibility to coronary heart disease and related diseases is strong, as is the evidence that monozygotic (genetically identical) twins tend to have more similar life-spans than do like sex dizygotic (genetically different, fraternal) twins. (G.A.Sa.)

SENESCENCE IN PLANTS

The growth of the vascular plant depends upon the activity of meristems, which are, in a sense, always embryonic. Continued indefinitely, this mode of growth could mean immortality; indeed, the longest lived individual organisms ever to have existed on earth have been certain species of trees. Plants and plant parts, however, do die, and death is often not the consequence of accident or environmental stress but of physiological decline—aging, or senescence.

Various kinds of physiological senescence and death occur and may affect particular cells, tissues, organs, or the whole plant. In the formation of the vessels of the xylem, cells conclude their differentiation by dying and contribute their empty walls to the conducting tissue. Individual organs such as leaves usually have a limited life-span. Entire shoot systems may gradually die back in the aerial parts of perennial plants, which overwinter underground. And, finally, the whole plant may die after a limited period of growth and the completion of reproduction. This behaviour is found in many annual plants, which complete their life cycle in a single growing season. The life-span may extend to two years, as in biennial plants, or longer, as in banana and certain bamboos, which die after flowering and fruiting.

Pro-grammed death of plant cells

In the examples cited above, the death of cells, organs, or individual plants appears to be "programmed" and, in some sense, adaptive. This is clearly so with the death of individual cells during differentiation, when residual products contribute to the effective function of the entire plant body. The death of leaves and of shoot systems is part of the plant's adaptation to the cycle of the seasons. In annual species, the death of the whole individual may be viewed in a similar way. The succession of generations in this case is carried on by seeds; the sacrifice of the parent plant may, in fact, contribute to the success of the seedling by making available to the seed a pool of reserves derived from the breakdown of parent tissues.

Certain features characterize the onset of senescence. The cells show degenerative changes often associated with the accumulation of breakdown products. Metabolic changes accompany the degeneration. Respiration may increase for a period, but the rate ultimately declines as the cellular apparatus degenerates. Synthesis of proteins and nucleic acids ceases, and, in some instances, disintegration of cells has been associated with the release of enzymes through the disruption of membrane-bounded bodies called lysosomes.

The death of individual cells in tissues such as the xylem appears to be governed by internal factors, but senescence often depends upon interaction of tissues and organs. The presence of young developing leaves often accelerates the aging of older leaves; removal of the younger leaves retards the senescence of the older ones, suggesting control by competition for nutrients. A similar effect is seen in annual plants, in which the development of fruits and seeds is associated with the senescence and, ultimately, the death of the rest of the plant; the removal of reproductive structures slows the rate of aging. In these instances competition obviously has some effect, but it does not sufficiently explain why older, mature organs suffer in competition with those still in active development. The link may lie partly in the capacity of developing organs to draw nutrients to themselves, even from older parts of the plant. Developing organs thus provide "sinks" toward which nutrients tend to move. The senescence of organs drained in this way could result from the progressive loss of certain key constituents; should leaf protein, for example, turn over by breakdown of proteins to their amino-acid constituents and then be resynthesized, a steady drain of amino acids from the leaf would progressively deplete the proteins in the leaf.

Nutritional "sinks"

"Sinks" can be only part of the explanation, however, for in detached leaves of plants such as tobacco, protein synthesis decreases, and protein content falls, while the amino-acid content actually rises. Senescence in such instances can hardly depend on the withdrawal of nutrients. Furthermore, leaf senescence can be retarded locally by the

application of cytokinins, hormones that stimulate plant cell division. Parallel effects have been demonstrated with growth substances of the auxin type in other plant systems. In the same way that active buds and fruits form sinks for nutrients from elsewhere in the plant, a cytokinin-treated area of a leaf attracts nutrients from other parts of the leaf. Although the metabolism of isolated leaves may differ in many respects from that of attached leaves, leaf senescence probably does not result only from nutrient drainage but also from the synthetic activity of leaf tissues, which may be under hormonal control from other parts of the plant. The root may be important, for roots are known to export cytokinins to the shoot.

Daily length of darkness

Environmental factors, primarily photoperiod (daily length of darkness) and temperature, play important parts in governing senescence and death in plants. In annual plants, death is the natural conclusion of development; thus, conditions accelerating development automatically advance senescence. This is readily seen in short-day plants, in which precocious reproduction upon exposure to long dark periods is followed by early death. Senescence may be retarded in these cases, however, by hormonal treatments of the kind known to delay degeneration and death in detached leaves. Competition for nutrients between vegetative and reproductive structures cannot be the primary cause of death, for, in species such as hemp, the male plants—which do not produce seeds—die earlier than the females under short-day (long-night) conditions.

In perennial plants, leaf fall is associated with approaching winter dormancy. In many trees leaf senescence is brought about by declining day length and falling temperature toward the end of the growing season. Chlorophyll, the green pigment in plants, is lost; yellow and orange pigments called carotenoids become more conspicuous; and, in some species, anthocyanin pigments accumulate. These changes are responsible for the autumn colours of leaves. There are some indications that day length may control leaf senescence in deciduous trees through its effect on hormone metabolism, for both gibberellins and auxins have been shown to retard leaf fall and to preserve the greenness of leaves under the short-day conditions of autumn.

From the foregoing it may be seen that senescence and death are important in the general economy of plants. The paradox that death contributes to survival is resolved when it is understood that the death of the part contributes to the better adaptation of the whole—whether organ, individual, or species. Viewed in this way, death is no more than another—albeit the ultimate—manifestation of development.

(J.H.-H.)

SENESCENCE IN MAMMALS

Changes in body composition, metabolism, and activity. The lean body mass, consisting of the skeletal muscles and all other cellular tissues, decreases steadily after physical maturity until, in extreme old age, it may be reduced to two-thirds its value in young adults. Body weight, however, usually increases with age, because stored fat and body water increases in excess of the loss of lean body mass. The relative amount of extracellular fluid increases with age during adult life, after decreasing steadily throughout fetal and postnatal development. Despite appearances, therefore, all tissues, even the skin, become more laden with water as a consequence of aging. The steady loss of voluntary (striated) muscle tissue mass throughout adult life depends somewhat on the pattern of physical activity. Evidence indicates that a large part of the loss of muscle mass with age is the result of disuse and atrophy rather than loss of muscle fibres.

The decrease of lean body mass is accompanied by a decrease in the level of overall metabolic activity. Basal metabolism is greatest during the period of most rapid mass growth; it then declines rapidly until physical maturity is reached and more slowly thereafter. In the rat the slow phase of decrease amounts to about 20 percent over a three-year period. The interior body temperature is maintained, despite lower heat production, by decreased blood flow through the skin with a consequent decrease of heat loss; the "cooling of the blood" with age, therefore, does not occur in the degree that might be inferred from the decrease in skin temperature. The amount of voluntary physical activity, such as running in an exercise wheel, typically decreases with age but varies considerably between individual animals.

Changes in structural tissues. The structural integrity of the vertebrate organism depends on two kinds of fibrous protein molecules, collagen and elastin. Collagen, which constitutes almost one-third of the body protein, is found in skin, bone, and tendons. When first synthesized by cells called fibroblasts, collagen is in a fragile and soluble form (tropocollagen). In time this soluble collagen changes to a more stable, insoluble form that can persist in tissues for most of an animal's life. The rate of collagen synthesis is high in youth and declines throughout life, so that the ratio of insoluble to soluble collagen increases with age. Insoluble collagen then builds up with age as a result of synthesis exceeding removal, much like another fibrous tissue, the crystalline lens of the eye. With increasing age, the number of cross-linkages within and between collagen molecules increases, leading to crystallinity and rigidity, which are reflected in a general body stiffness. There is also a decrease in the relative amount of a mucopolysaccharide (*i.e.,* the combination of a protein and a carbohydrate) ground substance; a measure of this, the hexosamine–collagen ratio, has been investigated as an index of individual differences in the rate of aging. An important consequence of these changes is decreased permeability of the tissues to dissolved nutrients, hormones, and antibody molecules.

The rate of aging of collagen is related to the overall metabolic activity of the animal; rats kept on low calorie diets have more youthful collagen than fully nourished rats of the same age.

Elastin is the molecule responsible for the elasticity of blood vessel walls. With age, progressive loss of elasticity of vessels occurs, presumably because of fragmentation of the elastin molecule.

Chemical retardation of aging

The cross-linkage of collagen is chemically similar to the cross-linkages that occur in skins when they are tanned to leather. This similarity has stimulated proposals that chemicals that inhibit cross-linkage in tanning will retard aging. Such compounds have been tried on animals, but the problems of toxicity have not yet been solved.

Tissue cell loss and replacement. The tissues of the body fall into two groups, according to whether or not there is continuous renewal of tissue cells. At one extreme are nonrenewal tissues such as nerves and voluntary muscles, in which no new cells are formed (at least in mammals) after a certain stage of growth. In renewal tissues such as the intestinal epithelium and the blood, on the other hand, some cell types live only one or a few days and must be replaced hundreds of times in the life-span of even a short-lived animal such as the rat. Between these limits lie many organs, such as liver, skin, and endocrine organs, that have cells that are replaced over periods ranging from a few weeks to several years in man.

A peripheral nerve is a convenient object to study because the total number of fibres in the nerve trunk can be counted. This has been done for the cervical and thoracic spinal nerve roots of the rat, the cat, and man. In the ventral and dorsal spinal roots of man, the number of nerve fibres decreases about 20 percent from age 30 to age 90. In the cat, the rat, and the mouse, however, the data do not consistently indicate a decrease of number of spinal root fibres with age. In man the number of olfactory nerve fibres, which serve the sense of smell, decreases by age 90 to about 25 percent of the number present at birth, and the number of optic nerve fibres, serving vision, decreases at a nearly comparable rate.

There is a striking decrease in the number of living cells in the cerebral cortex of the brain of humans with age. The cerebellar cortex of the rat and man is about as susceptible to age deterioration as is the cerebral cortex. Other parts of the brain are not so obviously marked by aging.

There is, in short, a tendency for the higher and more recently evolved levels of the nervous system to undergo more severe aging loss than do other regions, such as the brain stem and spinal cord. It is not yet known how much

of the loss of brain cells results from conditions within the brain itself and how much results from extrinsic causes, such as deterioration of the blood circulation. The nutrition and maintenance of nerve cells, or neurons, in the central nervous system depends to a considerable extent on neuroglia, small cells that surround the neurons. The absolute number of these cells apparently does not decrease with age, but some of the microscopic changes seen in the neurons of old persons are similar to the changes produced by starvation or physical exhaustion.

It has been shown that after an attack of measles, the virus remains in the host's body for the remainder of life and infrequently gives rise to a rapidly progressing degeneration of the cerebral cortex. This virus or other inapparent viruses may also be responsible for the individual differences in onset of senility in man.

The renewal tissues are typically made up of a population of proliferative cells, which retain the capability for division, and a population of mature cells, produced by the proliferative cells and with limited life-spans. The production of cells must balance the steady loss and also compensate quickly for unusual losses caused by injury or disease, so each renewal tissue has one or more channels of feedback control to adjust production to demand. Aging of renewal tissues is expressed in several ways, including decrease in the number of proliferative cells, decrease in the rate of cell division, and decrease in responsiveness to feedback signals. Changes of these factors in the blood-forming tissues of the mouse are small, yet the blood-forming tissues do suffer an aging deficit, for the ability to respond to extreme or repeated demand is significantly reduced in older mice.

Senescence of renewal tissues

The intact skin has a cell turnover time of several weeks, with the capability, shared by all renewal tissues, of temporarily increasing the rate of cell production by a large factor in response to injury. The rate of wound healing decreases with age, rapidly at first and more slowly as age increases.

One of the most regular and striking aging processes is the decrease in the ability to focus on both close and distant objects. This loss in visual accommodation is the result in part of a weakening of the ciliary muscle of the eye and of a decrease in the flexibility of the lens. A further contributing factor, however, is that the lens continues to grow throughout life at a rate that diminishes with age. This growth is the result of continuous division of epithelial cells near an imaginary midline of the lens, giving rise to fresh cells that differentiate into the precisely aligned lens fibres. Once formed, the fibres remain permanently in place.

An important feature of the renewal mechanism is the stem cell. These cells, which may normally continue to divide at a low rate throughout life, under conditions of increased demand enter a compensatory proliferative phase during which they divide rapidly. Blood-forming tissue has a stem-cell population that responds to injury readily in youth, but its capacity diminishes with age. The increased incidence of anemia in old age and the reduced capacity to respond to blood loss have been attributed to depletion of the blood-forming stem cells. Stem-cell populations have not been identified with certainty in other proliferative tissues. The intestinal mucosa, in particular, has a high cell-division rate without any clear indication of a reserve population of stem cells.

Mammalian cell cultures. Dividing cells from various mammalian tissues can be grown in vitro (outside the body) under careful laboratory control. Various lines of cancer cells have been grown in continuous culture for many decades. In the early period of tissue-culture technology it was claimed that certain chicken cells (fibroblasts) had been maintained in culture for 20 years. This led to the belief that dividing cells were potentially immortal and focussed interest on nondividing cells as the seat of the aging process. This view has lost standing in recent years. It has now been established that a population (clone) of fibroblasts has a finite life history in culture. It has a period of healthy growth, during which it can be transferred, or "split," several dozen times, indicating that the cells have undergone more than that number of generations. The cultures, however, go into a senescent phase and die out, usually before the 50th transfer. Occasionally, the chromosomes in a cell in the culture undergo a mutation (change) that results in a loss of a growth-limiting factor, leading to the establishment of a subclone capable of indefinite growth. This happens fairly often in cultures of mouse cell strains but only rarely in cultures of human cells. Such mutations usually involve chromosomal rearrangements or changes in the number of chromosomes.

The present view, therefore, is that dividing mammalian cells with a normal chromosomal complement have a limited growth potential and that the capacity for indefinite growth shown by cancer cells and transformed cells is the result of the loss of a growth-limiting factor. The number of transfers that cell strains can undergo decreases as the age of the donor increases, in a way reminiscent of the decreased turnover rate of fibroblasts in living chickens and of the decreased rate of wound healing with age.

Changes in tissue and cell morphology. There are numerous instances of tissue changes with age. The atrophy of tissues of moderate degree is usual. The shrinkage of the thymus is especially striking and important in view of its role in immunological defense. The diminution of cellular tissue and replacement by fatty or connective tissue is prominent in marrow and skin. In the kidney, entire secretory structures (nephrons) are lost. The secretory cells of the pancreas, thyroid, and similar organs decrease in numbers.

Wasting of tissues with age

An important age change is the accumulation of pigments and inert—possibly deleterious—materials within and between cells. The pigment lipofuscin accumulates within heart muscle cells; it is not detectable at ten years of age but rises to almost 3 percent of the cell volume by age 90. Amyloid substance, a protein–carbohydrate complex, increases in tissues in middle age; it is presumably a product of autoimmune reactions, immune reactions misdirected against the organism itself. In an extreme case of a rare autoimmune disease, amyloid disease, particular organs are virtually choked with amyloid substance. Trace metals also accumulate in various tissues with age, and although the amounts are very small, certain metals can poison enzyme systems, stimulate mutations, or cause cancer.

AGING AT THE MOLECULAR AND CELLULAR LEVELS

Aging of genetic information systems. The physical basis of aging is either the cumulative loss and disorganization of important large molecules (*e.g.*, proteins and nucleic acids) of the body or the accumulation of abnormal products in cells or tissues. A major effort in aging research has been focussed on two objectives: to characterize the molecular disruptions of aging and to determine if one particular kind is primarily responsible for the observed rate and course of senescence; and to identify the chemical or physical reactions responsible for the age-related degradation of large molecules that have either informational or structural roles. The working molecules of the body, such as enzymes and contractile proteins, which have short turnover times, are not thought to be sites of primary aging damage. The deoxyribonucleic acid (DNA) molecules of the chromosomes appear to be potential sites of primary damage, because damage to DNA corrupts the genetic message on which the development and function of the organism depend. Damage at a single point in the DNA molecule can be followed by the synthesis of an incorrect protein molecule, which may result in the malfunction or death of the host cell or even of the entire organism. Attention therefore has been given to the somatic mutation hypothesis, which asserts that aging is the result of an accumulation of mutations in the DNA of somatic (body) cells. Aneuploidy, the occurrence of cells with more or less than the correct (euploid) complement of chromosomes, is especially common. The frequency of aneuploid cells in human females increases from 3 percent at age 10 to 13 percent at age 70. Each DNA molecule consists of two complementary strands coiled around each other in a double helix configuration. Evidence indicates that breaks of the individual strands occur with a higher frequency than was once suspected and that virtually all such breaks are repaired by an enzymatic mechanism that

DNA damage

destroys the damaged region and then resynthesizes the excised portion, using the corresponding segment of the complementary strand as a model. The mutation rate for a species is therefore governed more by the competence of its repair mechanism than by the rate at which breaks occur. This may help to explain why the mutation rates of different species are roughly proportional to their generation times and justifies research to determine whether the enzymatic mechanisms involved are accessible to control. It remains to be seen whether a reduction of mutation rates will retard the onset of generalized aging or of a specific disease process.

There are, however, serious objections to the somatic mutation theory. The wasp *Habrobracon* is an insect that reproduces parthenogenetically (*i.e.*, without the need of sperm to fertilize the egg). It is possible to obtain individuals with either a diploid, or paired, set of chromosomes, as in most higher organisms, or a haploid, single, set. Any gene mutation in a haploid cell at an essential position would result in loss of a vital process and impairment or death of the cell; in a diploid cell a serious mutation is often compensated for by the complementary gene and the cell can carry on its vital functions. Experiments have shown that haploid wasps live about as long as diploids, implying either that mutations are not a quantitatively important factor in aging or that parthenogenetic species have compensated for the vulnerability of their haploids by developing an increased effectiveness of DNA repair.

Chromosomes can be separated into DNA and protein molecules, but with increasing difficulty in older cells. The isolated DNA of old animals, however, does not differ from that of the young. Although most of the DNA in a given cell at a given time is repressed (*i.e.*, blocked from functioning), it is more repressed in old animals; it is not yet known whether this is a primary age change or a consequence of reduced cell metabolism arising from other causes.

Aging of the immunological system. Another important molecular information system of the body is the immunological system, part of which, the thymus-dependent subsystem, is specialized for defense against invading microorganisms and for the as yet poorly understood role of detecting and removing body cells that have changed in such ways that they are no longer recognized by the body as part of its own substance, leading to the autoimmune reactions mentioned above. The immunological system has been implicated in the body's defenses against cancer. Cancerous growths (neoplasms) are thought to arise from single cells that undergo a drastic transformation as a result of either a genetic mutation or the activation of a latent (hidden) virus that may have been transmitted genetically from parent to offspring. The control of cancer susceptibility by genetically governed defense mechanisms has been indicated by the breeding of high and low cancer susceptibility in mice. There is a growing body of evidence that the thymus-dependent immunological system is instrumental in repressing the development of cancer.

Aging and cancer

One piece of evidence is that the immunosuppressive procedures of organ transplantation are often followed by a greatly increased incidence of neoplasms. The thymus-dependent system can itself, however, give rise to age-related autoimmune disease, in which the immunological system perceives normal body tissue as foreign and attacks it with antibodies. The initial step in these diseases is considered to be a somatic mutation in a single cell of the immunological system. Such considerations are the basis of several immunological theories of aging, which seek to explain the phenomena of senescence in terms of mutations in the immunological system.

Aging of neural and endocrine systems. The loss of psychological and neurophysiological capacities with age is undoubtedly the result, in large part, of the loss of neurons, but deficiencies in the metabolic processes of the surviving cells are demonstrably involved. The ability of the eye to dark-adapt (*i.e.*, increase its sensitivity at low light levels) decreases with age, but part of that decrease can be restored by breathing pure oxygen. Various mental processes in old people are also found to be improved by breathing oxygen. The establishment of a memory trace (connections in the brain that are associated with memory) involves the synthesis of protein; any slowed induction of protein synthesis, as from lower oxygen intake, with age could be a factor in the deficits of learning and memory of old people.

A general characteristic of aging of the endocrine system is that the cells that once responded vigorously to hormones become less responsive. A normal chemical in cells, cyclic adenosine monophosphate (AMP), is thought to be a transmitter of hormonal information across cell membrane; it may be possible to identify the specific sites in the membrane or the cell interior at which communication breaks down.

INTERNAL AND EXTERNAL CAUSES OF AGING

External environmental agents. *Ionizing radiations.* The shortening of life caused by ionizing radiations (*e.g.*, X-rays) has been determined for many species, including mice, rats, hamsters, guinea pigs, and dogs. The occurrence of some diseases, such as leukemia, may increase disproportionately after irradiation, with the degree of increase influenced by age and sex.

The permanent nature of radiation damage is shown by the comparison of life-spans of irradiated and control populations. An irradiated population dies out like a chronologically older unirradiated population. Members of a population given a single dose of X-rays or gamma rays in early adult life die of the same diseases that afflict the unirradiated control population, but they die months or even years earlier.

Continuous irradiation throughout life at low dose rates (daily doses from one-thousandth to one-tenth the dose that would kill immediately) speeds the mortality process. It is not yet clear if the molecular damage produced by such irradiation is the same as the molecular changes that accompany natural aging. Studies of animals and of cells grown in culture suggest that large doses of radiation kill by producing deleterious rearrangements of chromosomes in the proliferative cell population. Such aberrations also increase with age, but they seem to be less important in the natural aging process. At low radiation doses, chromosome aberrations become relatively less important than other effects, and the primary radiation damage in these conditions may bear a closer relation to the aging lesion. Under conditions of low-dose irradiation, however, the only definite effect is slightly increased cancer incidence; a generalized aging effect has not yet been observed.

The aging effect of natural radiation

Natural radioactivity in the body, arising mostly from radioactive potassium and radium, and natural background irradiation, from the Earth and from cosmic rays, are not major contributors to the aging process, even in the long-lived human species. They are responsible, however, for a small percentage of cancer incidence. Although the dose to the body from medical radiations is a fraction of the background level and the radiation from nuclear weapon test fallout is less than 1 percent of the background, both sources contribute to cancer induction in proportion to their amounts.

Temperature. Flour beetles, fruit flies, fishes, and other poikilothermic (temperature-variable) organisms live longer at the lower range of environmental temperature. These observations led to the rate-of-living hypothesis, which, simply stated, holds that an organism's life-span is dependent on some critical substance that is exhausted more rapidly at higher temperature. Careful analysis of the data on temperature–longevity relations shows, however, that the rate-of-living hypothesis is inadequate in its original form. The most telling evidence comes from experiments in which fruit flies were kept at one temperature for part of their lives and at another temperature for the remainder. The results are not consistent with the rate-of-living hypothesis, but no satisfactory theory has appeared as yet to take its place. An important factor that has not yet been adequately taken into account is the relation of metabolic efficiency to temperature. The energy cost of the biosynthetic processes studied has been discovered to be minimal at an intermediate temperature in the range to which the species is adapted and to increase at higher or lower temperatures. A related phenomenon holds for

longevity; the number of calories expended by fruit flies per lifetime is maximal at an intermediate temperature, so the rate of aging per calorie is minimal at that temperature.

There is a question of the degree to which aging occurs as a result of heat destruction (thermal denaturation) of proteins. Thermal denaturation is predominately a disruption of the folding of molecules, which requires the breaking of numbers of low-energy bonds. It seems not to be a strong contributing factor to aging. There is still the possibility that rare events, such as mutations, may arise to a significant degree from thermal denaturation.

Physical wear of nonrenewable structures. One of an animal's most important assets is its chewing apparatus, including jaws and teeth. Adaptation to tooth rate of wear is especially important for animals that consume large quantities of grass and herbage. Such adaptations include higher tooth crowns (hypsodonty), larger grinding area, and longer tooth growth period. Tooth wear may be limiting for survival in adverse environments, but, on the whole, it is not an important life-limiting characteristic. The same can be said for other external organs subject to physical wear.

Infectious disease and nutrition. The populations in poor environments, characterized by high rates of infectious disease and poor nutrition, have higher death rates than populations in good environments at all ages, yet there is no positive evidence that disadvantaged populations experience a higher rate of aging.

Rats kept on diets restricted in calories live longer and have lower cancer incidence than do rats that are allowed to eat at will; maximum longevity, however, is achieved at a nutritional level that keeps the animal sexually immature and below normal weight.

Internal environment: consequences of metabolism. The metabolic activities of organisms produce highly reactive chemicals, including strong oxidizing agents. The internal structure of the cell, however, minimizes the harmful effects of such agents; the critical reactions take place within enclosed structures such as ribosomes, membranes, or mitochondria, and counteractive enzymes such as peroxidases are present in abundance. It is nevertheless likely that low concentrations of these reactive substances can reach vital molecules and contribute to the characteristic rate of aging injury. Experiments in which mice are fed low levels of antioxidants such as butylated hydroxytoluene (BHT) have been encouraging but are still somewhat equivocal.

Aging and membrane integrity

Membranes are the site of much of the metabolic activity of cells; they provide the barriers that keep incompatible reactions separated. Certain membrane structures, called lysosomes, contain enzymes capable of digesting the cell if released; the stability of cells and organisms is therefore very much bound up with the stability of membranes. A number of drugs, including corticosteroids, salicylates, and antihistamines, act by stabilizing cell membranes against inflammatory stimuli. Some of them are found to prolong life in fruit flies and to prolong survival of cells in vitro. The mode of action of these drugs is connected to substances called prostaglandins, which can alter specific membrane characteristics. (G.A.Sa.)

Human aging

Aging begins as soon as adulthood is reached and is as much a part of human life as are infancy, childhood, and adolescence. Gerontology (the study of aging) is concerned primarily with the changes that occur between the attainment of maturity and the death of the individual. The goal of research in gerontology is to identify the factors that influence these changes. Application of this knowledge is expected to reduce the disabilities now associated with aging.

Human aging has many aspects, both for the individual and for society, but the primary ones fall into three major categories; namely, biological-physiological, psychological-behavioral, and social-economic.

The biological-physiological aspects of aging include both the basic biological factors that underlie aging and the general health status. Since the probability of death increases rapidly with advancing age, it is clear that changes must occur in the individual which make him more and more vulnerable to disease. For example, a young adult may rapidly recover from pneumonia, whereas an elderly person may die. (See above *Biological theories of aging.*)

Physiologists have found that the performance of many organs such as the heart, kidneys, brain, or lungs shows a gradual decline over the life-span. Part of this decline is due to a loss of cells from these organs, with resultant reduction in the reserve capacities of the individual. Furthermore, the cells remaining in the elderly individual may not perform as well as those in the young. Certain cellular enzymes may be less active, and thus more time may be required to carry out chemical reactions. Ultimately the cell may die.

EFFECT OF AGING OF THE BODY SYSTEMS

Heart disease

Cardiovascular system. Diseases of the heart are the single largest cause of death after age 65. Thus, with increasing age the heart becomes more vulnerable to disease. Even in the absence of detectable disease, the heart undergoes deleterious changes with advancing age. Structural changes include a gradual loss of muscle fibres with an infiltration of fat and connective tissue. There is a gradual accumulation of insoluble granular material (lipofuscin, or "age pigment") in cardiac muscle fibres. These granules, composed of protein and lipid (fat), make their first appearance by the age of 20 and increase gradually, so that by the age of 80 they may occupy as much as 5–10 percent of the volume of a muscle fibre.

The heart also shows a gradual reduction in performance with advancing age. The amount of blood pumped by the heart diminishes by about 50 percent between the ages of 20 and 90 years. There are marked individual differences in the effects of age. For example, some 80-year-old individuals may have cardiac function that is as good as that of the average 40-year-old individual.

Under resting conditions, the heart rate does not change significantly with age. During each beat, however, the muscle fibres of the heart do not contract as rapidly in the old as in the young. This reduction in power, or rate of work, is due to the age-associated reduction in the activities of certain cellular enzymes that produce the energy required for muscular contraction.

In spite of these changes, the heart, in the absence of disease, is able to meet the demands placed upon it. In response to physical exercise it can increase its rate to double or triple the amount of blood pumped each minute, although the maximum possible output falls, and the reserve capacity of the heart diminishes with age.

Blood vessels and blood pressure

Arteriosclerosis, or hardening of the arteries, increases markedly in incidence with age, and is often regarded as part of aging. This is not necessarily true. Arteriosclerosis may appear even in adolescents. It is a progressive disorder and is present to some extent in practically all individuals by middle life. It is, therefore, impossible to make a clear distinction between the effects of aging and the effects of disease in blood vessels in human beings. In some animal species, as, for example, the rat, that do not develop arteriosclerosis, age changes in the heart and blood vessels can be identified.

In general, blood vessels become less elastic with advancing age. There is a progressive thickening of the walls of larger blood vessels with an increase in connective tissue. The connective tissue itself becomes stiffer with increasing age. This occurs because of the formation of cross-links both within the molecules of collagen, a primary constituent of connective tissue, and between adjacent collagen fibres. These changes in blood vessels occur even in the absence of the deposits on the arterial wall characteristic of atherosclerosis, which interfere with blood flow through the arteries. The gradual loss of elasticity increases with resistance to the flow of blood so that blood pressure may increase. This in turn increases the work that the heart must do in order to maintain the flow of blood.

While both systolic and diastolic blood pressures (blood pressures at contraction and dilation of the heart, respectively) increase with age, the rate of systolic increase exceeds that of diastolic so that the pulse pressure widens (see Table 5). The increase in pressure stops in the eighth

Table 5: Mean Blood Pressure in Adults by Age and Sex, 1960–62
(millimetres of mercury)

age (years)	men		women	
	systolic	diastolic	systolic	diastolic
18–24	121.7	71.6	111.8	69.4
25–34	124.7	76.4	115.6	72.9
35–44	128.6	80.7	122.8	78.0
45–54	133.8	83.2	133.8	82.0
55–64	140.3	83.1	146.6	84.9
65–74	148.0	81.0	160.2	83.7
75–79	154.3	79.4	156.6	79.3

Source: U.S. Public Health Service.

decade of life, and there may even be a slight decline in pressure in extreme old age.

On the average, obese people have higher blood pressures than those with normal body weights. Since the incidence of obesity increases with age at least up to the age of 55–60, this factor may contribute in part to the increase in blood pressure with age.

Digestive system. Loss of teeth, which is often seen in elderly people, is more apt to be the result of long-term neglect than a result of aging itself. The loss of teeth and incidence of oral disease increase with age, but, as programs of water fluoridation are expanded and the incidence of tooth decay in children is reduced, subsequent generations of the elderly will undoubtedly have better teeth than the present generation.

While it is true that the secretion by the stomach of hydrochloric acid, as well as other digestive enzymes, decreases with age, the overall process of digestion is not significantly impaired in the elderly. Sugar, proteins, vitamins, and minerals are absorbed from the stomach and intestine as well in the elderly as in the young. Some investigations indicate a slight impairment in fat absorption, but the reduction is probably of little practical significance.

These findings have important implications for nutrition of the elderly. There is no evidence that the intake of any nutrient, such as vitamins and minerals, need be increased in the elderly because of impaired absorption. Nutritional deficiencies can be avoided as long as the diet is varied to assure adequate intake of all nutritional elements. Deficiencies are most likely to develop from poor eating habits, such as excessive intake of carbohydrate with a reduction in protein. In the elderly these deficiencies are most apt to be in the intake of protein, calcium, iron, vitamin A, and thiamine (also called vitamin B_1).

Nervous system. Changes in the structures of the brain due to normal aging are not striking. It is true that with advancing age there is a slight loss of neurons (nerve cells) in the brain. This is because, in the adult, neurons have lost the capacity to form new neurons by division. The basic number of neurons in the brain appears to be fixed by about the age of 10. The total number of neurons is extremely large, however, so that any losses probably have only a minor effect on behaviour. Since the physiological basis of memory is still unknown, it cannot be assumed that the loss of memory observed in elderly people is caused by the loss of neurons in the brain.

Neurons are extremely sensitive to oxygen deficiency. Consequently, it is probable that neuron loss, as well as other abnormalities observed in aging brains, results not from aging itself, but from disease, such as arteriosclerosis, that reduces the oxygen available to areas of the brain by reducing the blood supply.

Changes in the brain

There are probably functional changes in the brain that account for the slowing of responses and for the memory defects that are often seen in the elderly; and even small changes in the connections between cells of the brain could serve as the basis for marked behavioral changes, but, until more is known about how the brain works, behavioral changes cannot be related to physiological or structural changes. It is known that, because of the slow course of aging, the nervous system can compensate and maintain adequate function even in centenarians.

Human behaviour is highly dependent on the reception and integration of information derived from sensory organs, such as the eye and ear, as well as from nerve endings in skin, muscle, joints, and internal organs. There is, however, no direct relation between the sensitivity of receptors and the adequacy of behaviour, because the usual level of stimulation is considerably greater than the minimum required for stimulation of the sense organs. In addition, an individual adapts to gradual impairments in one sensory organ by using information available from other sense organs. Modern technology has also provided glasses and hearing aids to compensate for reduced acuity in the sense organs.

The incidence of gross sensory impairments, of which many are the result of disease processes, increases with age. One survey conducted in the United States classified 25.9 per 1,000 persons aged 65–74 as blind, in contrast to 1.3 per 1,000 aged 20–44 years. In the age group 65–74, 54.7 per 1,000 persons were classified as functionally deaf, compared with 5.0 per 1,000 in the age range 25–34 years.

Vision. Visual acuity (ability to discriminate fine detail) is relatively poor in young children and improves up to young adulthood. From about the middle 20s to the 50s there is a slight decline in visual acuity, and there is a somewhat accelerated decline thereafter. This decline is readily compensated for by the use of eyeglasses. There is also reduction in the size of the pupil with age. Consequently, vision in older people can be significantly improved by an increase in the level of illumination.

Aging also brings about a reduction in the ability to change the focus of the eye for viewing near and far objects (presbyopia), so that distant objects can ordinarily be seen more clearly than those close at hand. This change in vision is related to a gradual increase in rigidity of the lens of the eye that takes place primarily between the ages of 10 and 55 years. After age 55 there is little further change. Many people in their 50s adopt bifocal glasses to compensate for this physiological change.

The sensitivity of the eye under conditions of low illumination is less in the old than in the young; that is, "night vision" is reduced. Sensitivity to glare is also greater in the old than in the young.

The incidence of diseases of the eye, such as glaucoma and cataracts (characterized, respectively, by increased intra-ocular pressure and opaque lenses), increases with age, but recent advances in surgery and the development of contact lenses have made it possible to remove cataracts and restore vision to many individuals.

Hearing. Hearing does not change much with age for tones of frequencies usually encountered in daily life. Above the age of 50, however, there is a gradual reduction in the ability to perceive tones at higher frequencies. Few persons over the age of 65 can hear tones with a frequency of 10,000 cycles per second. This loss of perception of high frequencies interferes with identifying individuals by their voices and with understanding conversation in a group, but does not ordinarily represent a serious limitation to the individual in daily life. Listening habits and intellectual level play an important role in determining the ability to understand speech, so that there is often a disparity between measurements of pure tone thresholds and ability to perceive speech.

Loss of perception of high frequencies

Other sensory impairment. After the age of 70 other sense organs may show a reduction in sensitivity. Reduced taste sensitivity is associated with atrophy and loss of taste buds from the tongue in the elderly. The effect of aging on the sense of smell has not been precisely determined because this sense is extremely difficult to assess quantitatively; in addition, smoking and exposure to occupational odours and noxious substances in the air influence sensitivity to smells.

Sensitivity to pain is difficult to evaluate quantitatively under controlled laboratory conditions. There is some evidence that it diminishes slightly after the age of 70.

There is a general slowing of responses in the elderly. Reflexes become slightly more sluggish and the speed of conduction of impulses in nerves is slightly slowed. Old people require more time to respond to the appearance of a light than do young. The slowing with age is greater in situations where a decision must be made. For example,

more time is required to initiate a response in experiments in which the instructions are "Press the button with your right hand when the green light comes on, but with your left hand when the red light comes on" than if the instructions are, "Push the button if either light comes on." From these and other experiments it is concluded that the primary site of slowing of responses is within the brain rather than in the end organ (eye) itself.

Skin. The primary age change in the skin is a gradual loss of elasticity. Although this basic change plays a role, other factors, such as exposure to the weather and familial traits, also contribute to the development of wrinkles and the pigmentation associated with senescence. The ability of the skin to take up slack and remain closely adherent to the underlying structures is due to the presence of fibres of the proteins elastin and collagen. Studies of the minute structures of the skin show a gradual reduction in elastin. In addition, the collagen fibres show an increase in cross-links, which greatly restricts the elastic properties of the collagen network.

The effectiveness of facial massage in retarding the development of wrinkles has not been evaluated under carefully controlled conditions. The application of creams containing female sex hormone stimulates regeneration of skin and improves its elastic properties. Other effects, which may be undesirable or even hazardous, may follow repeated administration of these hormones.

Thyroid and adrenal hormones

Endocrine system. Because of the importance of hormones in the regulation of many physiological systems, impairments in endocrine (ductless) glands have traditionally been cited as important determinants in aging.

Thyroxine, the hormone secreted by the thyroid gland, regulates the level of activity of all the cells of the body. When thyroxine secretion is reduced, all metabolic processes proceed at a reduced rate and basal metabolism falls. (Metabolism consists of the chemical changes taking place within the cells of an organism during the processes of growth and restoration of tissues and the production of energy necessary for bodily processes; basal metabolism is the metabolism, as measured by the rate at which heat is given off, when an organism is in a resting and fasting state.) Since basal metabolism decreases with age, it seemed reasonable to ascribe aging to a loss of thyroid function, but this assumption has proved to be incorrect. Experimental studies have shown that the ability of the thyroid gland to produce thyroxine is not reduced in the elderly, and that there is a reduction in the utilization of thyroxine in various tissues of the body. Further studies of cellular metabolism are needed to find out why this is so.

Since aging is associated with reduced ability to adjust to stresses, and since the adrenal cortex (the outer part of the adrenal gland) plays a role in many of these adjustments, numerous attempts have been made to assess senescent changes in the function of the adrenal cortex. Although after the age of 50 there is a reduction in blood levels of the hormones secreted by the adrenal cortex, the ability of the gland to produce hormones when stimulated by the experimental administration of adrenocorticotrophic hormone (ACTH), the pituitary hormone that regulates the activity of the adrenal cortex, has been shown to be as good in the old as in the young.

Hormones of pituitary and of pancreas

The pituitary gland is often referred to as the master gland of the body, since it produces hormones that stimulate the activities of other endocrine glands, such as the adrenal, the thyroid, and the ovary. It was therefore once assumed that reduction in the function of these glands associated with aging is due to lack of proper stimulation from the pituitary gland. Methods for determination of the very small amounts of these regulating hormones present in the blood have been developed and as yet no systematic studies of age differences in blood levels of these hormones have been reported.

The pancreas secretes insulin, the hormone that regulates the utilization of sugar and other nutrients in the body. When the pancreas fails to produce adequate amounts of insulin, diabetes occurs. One test for diabetes involves measuring the rate of removal of sugar from the blood, that is, the glucose-tolerance test. One characteristic of aging is a reduction in the rate of removal of excess sugar from the blood. At present it is not known whether this represents the early stages of diabetes or whether it is a normal age change. It does appear in aged individuals who do not show any of the other symptoms of diabetes. Furthermore, it has been shown that, unlike the diabetic, elderly subjects can, with additional stimulation, produce more insulin. In normal young persons the pancreas releases more insulin in response to even a slight rise in blood sugar levels. In the elderly, the sensitivity of the pancreas is reduced so that a higher level of blood sugar is required to stimulate it to action. With maximum stimulation the pancreas in the aged can produce as much insulin as the pancreas in the young.

Sexual activity

It has long been known that the excretion of both male and female sex hormones diminishes with age. In the female, the excretion of estrogens (female sex hormones) falls markedly at the menopause (see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *The human reproductive system: The female reproductive system: Menopause*). In the male, the excretion of androgens (male sex hormones and their degradation products) falls gradually over the age span 50–90, so that the existence of a male "climacteric" is highly improbable.

Sexual activity, as reported in interview studies, diminishes progressively between the ages of 20 and 60 in both males and females. In males the frequency of marital intercourse falls from an average of four per week in 20-year-olds to one per week in 60-year-olds. Practically all males aged 20–45 reported some level of sexual activity. Between the ages of 45 and 60 only about 5 percent of males reported loss of sexual activity.

Few systematic studies have been made of sexual behaviour in individuals over the age of 60, but clinical reports indicate that at least some males remain sexually active at 90.

There are wide individual differences in the level of sexual activity in both males and females. In human beings, sexual behaviour is influenced more by psychological and social factors than by the levels of sex hormones circulating in the blood. Nevertheless, the use of male sex hormones has had a long, and stormy, history as a rejuvenating agent for males. Attempts to rejuvenate elderly males by injecting crude extracts from testicles of animals, as well as various androgenic compounds, were made, but the effects, if any, were only transitory. In the early 1900s, sex glands from other animals were transplanted into human beings, but the results were questionable and the side effects were often disastrous. At about the same time, an operation was devised in which the spermatic ducts were tied off. It was assumed that preventing the loss of sperm would stimulate the sex glands to produce androgenic hormones which would rejuvenate the individual. None of these assumptions proved correct, so that the operation was soon abandoned as a rejuvenating procedure.

Since tissue loss does occur with aging, the administration of anabolic hormones (hormones that promote the buildup of tissues) may represent an important future development. The compounds that are currently available have a number of undesirable side effects and cannot be used routinely. Chemists and pharmacologists continue research to produce new steroids that will have anabolic effects without the undesirable side effects.

Calcium loss

Skeletal system. With aging, the bones gradually lose calcium. As a result they become more fragile and are more likely to break, even with minor falls. Healing of fractures is also slower in the old than in the young. Recent advances in orthopedic surgery, with the replacement of parts of a broken bone or joint with new structures or the introduction of metallic pegs to hold broken parts together, have been of great value to elderly people.

The incidence of osteoporosis, a disease characterized by a loss of calcium and minerals from bone, also increases with age. It occurs more frequently in women after menopause than in men and is especially evident in the spinal column. Back pain is a primary symptom of the disease. It can be treated by increasing calcium intake in association with the administration of anabolic hormones.

The mobility of joints diminishes with age and the incidence of arthritis increases.

Respiratory system. Vital capacity, or the total amount of air that can be expelled from the lung after a maximum inspiration, diminishes with age, as does the total volume of air that can be contained in the lungs. In contrast, the amount of air that cannot be expelled from the lung increases. These changes in respiratory mechanisms are primarily a reflection of the increased stiffness of the bony cage of the chest and decreased strength of the muscles that move the chest during respiration.

The lung also contains elastin and collagen to give it elastic properties. As indicated previously, the formation of cross-links in elastin and collagen that takes place with aging reduces the elastic properties of the lung.

The transfer of oxygen and carbon dioxide from the air in the lungs to the blood is influenced by the amount of blood flowing through the lungs as well as by the amount of air moved in and out. The characteristics of the membranes that separate blood and air in the lungs are also important in maintaining an adequate supply of oxygen to the body. Although with age there is a slight reduction in the amount of oxygen that can be moved from the air to the blood in the lungs, the reduction becomes apparent only when large amounts of oxygen are required, as during strenuous exercise. It is believed that a primary factor in the impairment of oxygen transfer in the lungs of elderly subjects is the lack of appropriate adjustment of the blood flow to the air sacs in the lung.

Incidence of emphysema

Emphysema, abnormal distension of the lungs with air, is a lung disease reaching its highest incidence between the ages of 45 and 65. In the United States the death rate from emphysema increased by almost 400 percent between 1950 and 1960. Although the exact causes of the disease are still unknown, the presence of noxious or toxic agents in the air may be a contributing factor. Many studies have shown a relationship between the incidence of emphysema and bronchitis (inflammation of the bronchi) and smoking. Among British physicians death rates from bronchitis were six times higher in those smoking 25 cigarettes a day than in nonsmokers.

Effects of smoking

Measurements of lung function are significantly lower in cigarette smokers than in nonsmokers of the same age. Values for cigarette smokers are, on the average, about equal to those of nonsmokers who are 10–15 years older. There is evidence, however, that when cigarette smokers quit smoking, measurements of pulmonary function closely approach those of nonsmokers within one to two years, even in the case of heavy smokers 50–60 years old.

Kidney. The kidney removes wastes from the body by separating them from the blood and forming urine. In this process many substances are accumulated in the urine at a higher concentration than in the blood. With advancing age the concentrating ability of the kidney falls, so that a greater volume of water is required to excrete the same amount of waste material. This loss in concentrating ability is probably partially offset by a decrease in the excretory load because of reduced activity, alterations in food intake, and the reduction in muscle mass of the elderly. These changes in kidney function may not be reflected in urine volume, since volumes fluctuate widely at all ages and are determined primarily by fluid intake.

The reduction in renal (kidney) function is due in part to a gradual reduction in blood flow to the kidney. Since the kidney receives a great excess of blood (about 25 percent of the blood pumped by the heart each minute), the reduction with age does not normally result in an accumulation of waste products in the blood. Any such accumulation is the result of disease that damages the kidney. The reduced concentrating ability of the kidney results from a loss of some of the nephrons, the functional elements of the kidney, and the reduced activity of cellular enzymes.

Regulatory mechanisms. Some physiological characteristics, such as the mechanisms that regulate the acidity of the blood or its sugar level, are adequate to maintain normal levels under resting conditions even in very old people; however, the aged require more time than the young to reestablish normal levels when changes from the normal occur.

In order to test the effectiveness of control mechanisms of the body, physiologists produce changes experimentally and determine the rate of recovery. When the acidity of the blood is increased to the same extent in old and young subjects, it is returned to normal within 6–8 hours in the young; in the elderly 18–24 hours are required.

Similarly, the rate of return to fasting levels after sugar has been administered intravenously or orally is slower in the old than in the young. The response to insulin, which accelerates the removal of sugar from the blood, is also diminished in the elderly.

The body's physiological mechanisms for adjusting to changes in environmental temperature are less adequate in the old than in the young. Consequently older people may prefer more uniform and slightly higher temperatures than the young. High temperatures are also more hazardous to the elderly. The incidence of heat prostration in hot weather increases with age.

Exercise and work

Exercise is one of the physiological stresses of daily living. In reasonable amounts it is a valuable stimulus to maintain physiological vigour. A number of studies have indicated a lower incidence of cardiovascular disease among adults who indulge in physical activity than in those who do not.

The capacity to perform muscular work diminishes progressively in the elderly. Muscle strength diminishes; however, the reduction in strength is less in muscles that continue to be used throughout adult life than in those that are not. Thus a part of the reduction in muscle strength may be an atrophy of disuse.

Maximum work capacity is reduced in the elderly, largely because of the inability to deliver enough oxygen to the working muscles. In the young, the need for oxygen is met for the most part by increasing the heart rate. Under conditions of maximum work, young adults can increase their heart rate to over 200 beats per minute; the elderly to only about 150 per minute. In addition, the transfer of oxygen from the lungs to the blood is reduced in the elderly under conditions of strenuous exercise.

With less than maximum exercise, there is a greater increase in blood pressure, heart rate, and respiration in the old than in the young; that is, a given work load induces a greater physiological stress in the old than in the young. Furthermore, recovery of blood pressure, heart rate, and respiration to resting values takes longer in the old.

Premature aging. Progeria is an extremely rare disease of early childhood characterized by many of the superficial aspects of aging, such as baldness, thinning of the skin, prominence of blood vessels of the scalp, and vascular disease. These children have the general appearance of "little old men." They rarely live beyond the age of 15–18. Death is usually caused by cardiovascular disease.

The disease is extremely rare. In fact, only about 50 cases have been identified for study. In spite of the appearance of premature aging, these patients fail to show an acceleration of other age changes, except for the early development of cardiovascular disease. Most tests of physiological and psychological functions give values which are normal for their chronological age. It is doubtful whether the child with progeria is suffering only from accelerated aging; rather, progeria should be regarded as a rare disease with a superficial resemblance to senescence.

PSYCHOLOGICAL ASPECTS OF AGING

Intellectual functions in the aged

The most outstanding psychological features of aging are the impairment in short-term memory and the lengthening of response time. Both of these factors contribute to lower scores of the elderly on standard tests of "intelligence." When the aged are given all the time that they wish on tests that are not heavily dependent on school skills, their performance is only slightly poorer than that of young adults. Age decrements are negligible on tests that depend on vocabulary, general information, and well-practiced activities.

Experimental studies on learning show that, although the elderly learn more slowly than the young, they can acquire new material and can remember it as well as the young. Age differences in learning increase with the difficulty of the material to be learned.

Aged people tend to become more cautious and rigid in their behaviour and to withdraw from social contacts. These behaviour patterns may be the result of social

institutions and expectancies rather than an intrinsic phenomenon of aging. Many persons who "age successfully" make conscious efforts to maintain mental alertness by continued learning and by expansion of social contacts with individuals in a younger age group. (N.W.S.)

BIBLIOGRAPHY

Growth. General texts include SCOTT F. GILBERT, *Developmental Biology,* 4th ed. (1994), an attempt to integrate the new advances in molecular biology with cellular and organismal processes; KEITH ROBERTS *et al.* (eds.), *Molecular and Cellular Basis of Pattern Formation* (1991); J.D. MURRAY, *Mathematical Biology,* 2nd, corrected ed. (1993); N.K. WESSELLS and JANET L. HOPSON, *Biology* (1988); and LEON WEISS (ed.), *Histology: Cell and Tissue Biology,* 5th ed. (1983). D'ARCY WENTWORTH THOMPSON, *On Growth and Form,* 2nd ed. (1942, reissued 1992), also available in an abridged ed. edited by JOHN TYLER BONNER (1961, reissued 1992), is the classic mathematical treatment of the dynamics of growth. PETER L. ANTONELLI (ed.), *Mathematical Essays on Growth and the Emergence of Form* (1985), is a collection of articles. Other useful texts include A. HABENICHT (ed.), *Growth Factors, Differentiation Factors, and Cytokines* (1990); MICHAEL B. SPORN and ANITA B. ROBERTS (eds.), *Peptide Growth Factors and Their Receptors,* 2 vol. (1990); and MICHAEL J. REISS, *The Allometry of Growth and Reproduction* (1989).

Specific treatments of plant growth are JAMES D. MAUSETH, *Botany: An Introduction to Plant Biology,* 2nd ed. (1995); ARTHUR W. GALSTON, PETER J. DAVIES, and RUTH L. SATTER, *The Life of the Green Plant,* 3rd ed. (1980); KINGSLEY R. STERN, *Introductory Plant Biology,* 5th ed. (1991); and J.E. DALE and F.L. MILTHORPE (eds.), *The Growth and Functioning of Leaves* (1983). DENNIS R. CAMPION, GARY J. HAUSMAN, and ROY J. MARTIN (eds.), *Animal Growth Regulation* (1989), for advanced readers, explores growth regulation primarily in domesticated animals.

Malformations in the plant world are the topic of GEORGE N. AGRIOS, *Plant Pathology,* 3rd ed. (1988), a college-level text that presents the effect of pathogens on host growth and functioning in chapters 3–5; R. HEITEFUSS and P.H. WILLIAMS (eds.), *Physiological Plant Pathology* (1976); P.G. AYRES (ed.), *Effects of Disease on the Physiology of the Growing Plant* (1981), a compilation of seminar papers; WILLIAM F. BENNETT (ed.), *Nutrient Deficiencies & Toxicities in Crop Plants* (1993); and R.D. DURBIN (ed.), *Toxins in Plant Disease* (1981).

Animal malformations including human malformations are discussed in *Birth Defects Original Article Series* (monthly), authoritative articles on congenital malformations and abnormalities; MARY LOUISE BUYSE (ed.), *Birth Defects Encyclopedia* (1990); *Congenital Malformations Worldwide: A Report from the International Clearinghouse for Birth Defects Monitoring Systems* (1991); and ADAM S. WILKINS, *Genetic Analysis of Animal Development,* 2nd ed. (1993).

Biological regeneration is explored by CHARLES E. DINSMORE (ed.), *A History of Regeneration Research: Milestones in the Evolution of a Science* (1991); JOHN F. FALLON *et al.* (eds.), *Limb Development and Regeneration,* 2 vol. (1993); FREDERICK J. SEIL (ed.), *Neural Regeneration* (1994), highly technical symposium papers; JOHN G. NICHOLLS, *The Search for Connections: Studies of Regeneration in the Nervous System of the Leech* (1987), a short, highly readable monograph; HARRY J. BUNCKE (ed.), *Microsurgery: Transplantation—Replantation: An Atlas-Text* (1991); GERALD WEISSMANN (ed.), *The Cell Biology of Inflammation* (1980); DAVID EVERED and JULIE WHELAN (eds.), *Fibrosis* (1985), symposium papers; ROBERT P. MECHAM (ed.), *Regulation of Matrix Accumulation* (1986); ELIZABETH D. HAY (ed.), *Cell Biology of Extracellular Matrix,* 2nd ed. (1991); and ERKKI RUOSLAHTI and EVA ENGVALL (eds.), *Extracellular Matrix Components* (1994).

General features of biological development. A classic work that laid the foundation for the modern interpretation of development in terms of gene activities is THOMAS H. MORGAN, *Embryology and Genetics* (1934, reprinted 1975). Overviews of biological development include JONATHAN BARD, *Morphogenesis: The Cellular and Molecular Processes of Developmental Anatomy* (1990); LEON W. BROWDER (ed.), *The Cellular Basis of Morphogenesis* (1986), a comprehensive synthesis of developmental biology drawing upon knowledge from molecular biology, anatomy, and genetics; MERTON BERNFIELD (ed.), *Molecular Basis of Morphogenesis* (1993); V.E.A. RUSSO *et al.*, *Development: The Molecular Genetic Approach* (1992); and JAMES D. WATSON *et al.*, *Molecular Biology of the Gene,* 4th ed., 2 vol. (1987), for advanced undergraduate and graduate students. MONTGOMERY SLATKIN (ed.), *Exploring Evolutionary Biology: Readings from American Scientist* (1995), collects 31 essays and articles.

Studies of more specialized topics include RICHARD R. RIBCHESTER, *Molecule, Nerve, and Embryo* (1986), a college text; MARIT NILSEN-HAMILTON (ed.), *Growth Factors and Signal Transduction in Development* (1994), describing the interactions of growth factors and their receptors and the subsequent signal transduction pathways they activate in directing developmental processes; MICHAEL T. ZAVY and RODNEY D. GEISERT (eds.), *Embryonic Mortality in Domestic Species* (1994); and P.A. HAUSEN and METTA RIEBESELL, *The Early Development of Xenopus laevis: An Atlas of the Histology* (1991), on the key model (frogs) for developmental and embryological studies in vertebrates.

Plant development. JIŘÍ ŠEBÁNEK (ed.), *Plant Physiology,* trans. from Czech (1992), covers plant growth, development, and the resistance of plants to unfavourable abiotic and biotic effects. LINCOLN TAIZ and EDUARDO ZEIGER, *Plant Physiology* (1991), includes treatment of plant developmental processes; as does P.E. WAREING and I.D.J. PHILLIPS, *Growth and Differentiation in Plants,* 3rd ed. (1981). KALLIOPI A. ROUBELAKIS-ANGELAKIS and KIEM TRAN THANH VAN (THANH VAN KIEM TRAN) (eds.), *Morphogenesis in Plants: Molecular Approaches* (1993), presents advances in molecular and cellular biology pertaining to plant morphogenesis with particular reference to alternative approaches in solving difficulties of plant morphogenic expression. JIŘÍ ŠEBÁNEK, ZDENĚK SLADKÝ, and STANISLAV PROCHÁZKA (eds.), *Experimental Morphogenesis and Integration of Plants* (1991; originally published in Czech, 1983), addresses morphogenesis and structural integration in plants including possible ways of regulating these processes with regard to the practical needs of agriculture, horticulture, and silviculture.

Specific topics are treated in ROGER V. JEAN, *Mathematical Approach to Pattern and Form in Plant Growth* (1984; originally published in French, 1983); CLIVE W. LLOYD (ed.), *The Cytoskeleton in Plant Growth and Development* (1982); F.A.L. CLOWES, *Morphogenesis of the Shoot Apex* (1972); THOMAS A. STEEVES, *Endogenous Plant Growth Substances,* 2nd ed. (1980), a historical look at the discovery and significance of plant hormones; L.W. ROBERTS, P.B. GAHAN, and R. ALONI, *Vascular Differentiation and Plant Growth Regulators* (1988); THOMAS C. MOORE, *Biochemistry and Physiology of Plant Hormones,* 2nd ed. (1989); RICHARD N. ARTECA, *Plant Growth Substances: Principles and Applications* (1996); KARL J. NIKLAS, *Plant Allometry: The Scaling of Form and Process* (1994), the application of allometry to the studies of evolution, morphology, physiology, and reproduction of plants; ONKAR D. DHINGRA and JAMES B. SINCLAIR, *Basic Plant Pathology Methods,* 2nd ed. (1995), with more than 1,800 literature citations that can serve as research sources; B.M. JOHRI (ed.), *Embryology of Angiosperms* (1984); V. RAGHAVAN, *Embryogenesis in Angiosperms: A Developmental and Experimental Study* (1986); I.T. EVANS, *Daylength and the Flowering of Plants* (1975); RICHARD E. KENDRICK and BARRY FRANKLAND, *Phytochrome and Plant Growth,* 2nd ed. (1983); AUGUST DE HARTOGH and MARCEL LE NARD (eds.), *The Physiology of Flower Bulbs* (1993); and J. DEREK BEWLEY and MICHAEL BLACK, *Seeds: Physiology of Development and Germination,* 2nd ed. (1994), covering in chronological sequence the most important events in seed development and the maturation, germination, and postgermination stages of seedling establishment.

Animal development. General textbooks covering the subject include J. BRACHET and H. ALEXANDRE, *Introduction to Molecular Embryology,* 2nd totally rev. and enlarged ed. (1986); and GERALD M. EDELMAN, *Topobiology: An Introduction to Molecular Embryology* (1988). A history of modern experimental development is contained in the work by Gilbert, cited above. HANS SPEMANN, *Embryonic Development and Induction* (1938, reprinted 1988; originally published in German, 1936), is a classic exposition of the experimental method in embryology. Additional useful works are ROBERT WALL, *This Side Up: Spatial Determination in the Early Development of Animals* (1990); D.R. JOHNSON, *The Genetics of the Skeleton: Animal Models of Skeletal Development* (1986); JOHN PHILLIP TRINKAUS, *Cells into Organs: The Forces That Shape the Embryo,* 2nd ed. (1984); ELIZABETH S. WATTS (ed.), *Nonhuman Primate Models for Human Growth and Development* (1985); MATTHEW H. KAUFMAN, *The Atlas of Mouse Development* (1992); CLAUDIO D. STERN and PHIL W. INGHAM (eds.), *Gastrulation* (1992); BRIAN K. HALL, *The Neural Crest* (1988); BRIGID HOGAN *et al.*, *Manipulating the Mouse Embryo: A Laboratory Manual,* 2nd ed. (1994); and the work by Wilkins, cited above.

Human growth and development. Overviews of the topic are provided by MANUEL HERNÁNDEZ (M. HERNÁNDEZ RODRÍGUEZ) and JESÚS ARGENTE (eds.), *Human Growth: Basic and Clinical Aspects* (1992), conference proceedings; FRANK FALKNER and J.M. TANNER (eds.), *Human Growth,* 3 vol. (1978–79), with vol. 1 and 3 in a 2nd ed. (1986); BARRY BOGIN, *Patterns of Human Growth* (1988), a sophisticated discourse for advanced students and researchers in bioanthropology; and ESMAIL MEISAMI and PAOLA S. TIMIRAS (eds.), *Handbook of Human Growth and Developmental Biology,* 3 vol. in 7 (1988–90).

Explorations of human embryology include STEPHEN G.

GILBERT, *Pictorial Human Embryology* (1989), an atlas for the health professional and the general reader; KEITH L. MOORE and T.V.N. PERSAUD, *The Developing Human: Clinically Oriented Embryology,* 5th ed. (1993); JAN LANGMAN, *Langman's Medical Embryology,* 7th ed. by T.W. SADLER (1995), a concise presentation; and JAN S. ZAGON and THEODORE A. SLOTKIN (eds.), *Maternal Substance Abuse and the Developing Nervous System* (1992).

On human anatomy, ELAINE N. MARIEB, *Human Anatomy and Physiology,* 3rd ed. (1995); and ALEXANDER P. SPENCE, *Basic Human Anatomy,* 3rd ed. (1990), are widely used introductory texts for undergraduate students. Various aspects of human anatomy are covered in KEITH L. MOORE and ANNE M.R. AGUR, *Essential Clinical Anatomy* (1995); FRANK J. SLABY, SUSAN K. MCCUNE, and ROBERT W. SUMMERS, *Gross Anatomy in the Practice of Medicine* (1994); ANNE M.R. AGUR and MING J. LEE, *Grant's Atlas of Anatomy,* 9th ed. (1991); RONALD G. WOLFF, *Functional Chordate Anatomy* (1991), a textbook that integrates body functions across systems; KURT E. JOHNSON, *Human Developmental Anatomy* (1989); KENNETH M. BACKHOUSE and RALPH T. HUTCHINGS, *A Color Atlas of Surface Anatomy: Clinical and Applied* (1986); HENRY CLAY, *Anatomy of the Human Body,* 30th American ed. edited by CARMINE D. CLEMENTE (1985); JAMES E. CROUCH, *Functional Human Anatomy,* 4th ed. (1985); W. HENRY HOLINSHEAD, *Textbook of Anatomy,* 4th ed. (1985); and ROBERT L. BACON and NELSON R. NILES, *Medical Histology: A Text-Atlas with Introductory Pathology* (1983).

GILBERT B. FORBES, *Human Body Composition: Growth, Aging, Nutrition, and Activity* (1987), is a good source of introductory information for general readers. HAN C.G. KEMPER (ed.), *Growth, Health, and Fitness of Teenagers: Longitudinal Research in International Perspective* (1985), reports on a study of age-related changes in the physical growth and physical activity and performance of teenagers in The Netherlands. ROBERT M. MALINA and CLAUDE BOUCHARD, *Growth, Maturation, and Physical Activity* (1991), is a well-organized textbook for students interested in human growth and its relation to body composition and physical performance. R.J. SHEPHARD and J. PAŘÍZKOVÁ (eds.), *Human Growth, Physical Fitness, and Nutrition* (1991), contains a nicely presented, invaluable collection of research papers for health scientists and clinicians concerned with nutrition and physical fitness in children during their growing years. DAVID SINCLAIR, *Human Growth After Birth,* 5th ed. (1989), provides a clear, basic textbook for nursing and other health profession students and general readers.

Aging and senescence. *Life-span:* Studies of longevity include LEONID A. GAVRILOV and NATALIA S. GAVRILOVA, *The Biology of Life Span: A Quantitative Approach,* rev. and updated ed. (1991; originally published in Russian, 2nd ed., rev. and updated, 1991); and JOEP M.A. MUNNICHS *et al.* (eds.), *Life-Span and Change in Gerontological Perspective* (1985), a report of diverse research on behavioral development across the life-span. *United Nations Demographic Yearbook* includes mortality statistics for all countries, showing the influence of economic, social, and climatic factors on mortality.

Aging: General considerations are addressed by JOHN A. BEHNKE, CALEB B. FINCH, and GAIRDNER B. MOMENT (eds.), *The Biology of Aging* (1978), a collection of articles; MICHAEL R. ROSE, *Evolutionary Biology of Aging* (1991), a provocative treatise that proposes an explanatory theory of aging grounded in evolutionary biology; ROBERT R. KOHN, *Principles of Mammalian Aging,* 2nd ed. (1978), with emphasis on the role of interstitial tissue changes in the aging process; BERNARD L. STREHLER, *Time, Cells, and Aging,* 2nd ed. (1977), an examination of cellular and molecular mechanisms and theories of aging; BRIAN CHARLESWORTH, *Evolution in Age-Structured Populations,* 2nd ed. (1994), a theoretical consideration of the consequences of age-structure and age-specific differences in reproduction and mortality, which also considers the broader issue of life-history evolution and hence treats senescence as a part of the continuum of development; CALEB E. FINCH, *Longevity, Senescence, and the Genome* (1990), an encyclopaedic treatment of the theories of aging, the statistical methods for evaluating aging, and the full range of empirical methods used to study the aging process for all levels of biological organization, ranging from molecules to populations, with tables and summaries of observed maximum life spans and mortality rates; and ROBERT E. RICKLEFS and CALEB E. FINCH, *Aging: A Natural History* (1995).

Human aging: GEORGE L. MADDOX *et al.* (eds.), *The Encyclopedia of Aging,* 2nd ed. (1995), is an extensive general reference. General principles are addressed in PAOLA S. TIMIRAS (ed.), *Physiological Basis of Aging and Geriatrics,* 2nd ed. (1994); IMRE ZS.-NAGY, *The Membrane Hypothesis of Aging* (1994), a comprehensive, multidisciplinary description of the cell maturation and aging process; ARTHUR K. LABIN (ed.), *Practical Handbook of Human Biologic Age Determination* (1994), covering metabolic profiles, organ system approaches, biological measurements, and nonhuman model studies; ROBERT ARKING, *Biology of Aging: Observations and Principles* (1991); and ALEXANDER P. SPENCE, *Biology of Human Aging* (1989).

The following provide advanced knowledge on various aspects of geriatrics and gerontology: JEAN M. LAUDER *et al.* (eds.), *Molecular Aspects of Development and Aging of the Nervous System* (1989); CLAIRE MURPHY, WILLIAM S. CAIN, and D. MARK HEGSTED (eds.), *Nutrition and the Chemical Senses in Aging: Recent Advances and Current Research Needs* (1989); ALLAN L. GOLDSTEIN (ed.), *Biomedical Advances in Aging* (1990); and STEVEN R. GAMBERT (ed.), *Handbook of Geriatrics* (1987). The essentials of geriatric medicine are presented in WILLIAM REICHEL, *Care of the Elderly: Clinical Aspects of Aging,* 4th ed. (1995), a comprehensive clinical approach geared to the practicing physician; M.R.P. HALL, W.J. MACLENNAN, and M.D.W. LYE, *Medical Care of the Elderly,* 3rd ed. (1993); and *Contemporary Geriatric Medicine* (biennial).

Additional works include ALAN J. SINCLAIR and KEN W. WOODHOUSE (eds.), *Acute Medical Illness in Old Age* (1995); RICHARD L. BYYNY and LEON SEROFF, *A Clinical Guide for the Care of Older Women: Primary and Preventive Care,* 2nd ed. (1996); CORNELIUS L.E. KATONA, *Depression in Old Age* (1994); MARGRET M. BALTES and PAUL B. BALTES (eds.), *The Psychology of Control and Aging* (1986); and GERALD FELSENTHAL, SUSAN J. GARRISON, and FRANZ U. STEINBERG (ed.), *Rehabilitation of the Aging and Elderly Patient* (1994). (Ed.)

Guyana

Guyana (the Co-operative Republic of Guyana) is an independent republic and member of the Commonwealth located in the northeastern corner of South America. It is bordered by Venezuela to the west, Brazil to the southwest and south, Suriname (along the Courantyne River) to the east, and the Atlantic Ocean to the north. Its total area of 83,000 square miles (215,000 square kilometres) is largely uninhabited, and most of the country's inhabitants occupy the narrow coastal strip. The capital and chief port is Georgetown.

Present-day Guyana reflects its British colonial past and its reactions to that past. It is the only English-speaking country of South America. Since independence in 1966, Guyana's chief economic assets—its sugarcane plantations and bauxite industry—have come under government control, as has most of the country's commerce. Guyana's populace is mainly of colonial origin, although a small number of aboriginal Indians are scattered throughout the forested interior.

The more numerous coastal peoples are chiefly descendants of slaves from Africa and indentured workers from India, who were originally imported to work the coastal sugarcane plantations. Racial problems between the latter two groups have played a disruptive role in Guyanaese society.

Politically, Guyana has moved on a steady course toward socialism from the time of independence, although after the death of the first prime minister, Forbes Burnham, in 1985, ties with Western powers were strengthened.

This article is divided into the following sections:

Physical and human geography 441
- The land 441
  - Relief
  - Drainage
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 443
- The economy 443
  - Resources
  - Agriculture, forestry, and fishing
  - Industry
  - Finance and trade
  - Transportation
- Government and social conditions 445
  - Government
  - Education
  - Health and welfare
- Cultural life 445

History 445

Bibliography 446

Physical and human geography

THE LAND

Relief. The narrow plain that extends along the country's Atlantic coast has been modified considerably by humans. Much of the area, which measures only about 10 miles (16 kilometres) at its widest point, has been reclaimed from the sea by a series of canals and some 140 miles of dikes. The coastal plain's inland border is generally marked by canals that separate the plain from interior swamps. South of the coastal zone the forested land rises gently and has sandy soils.

The coastal regions

About 40 miles inland from the coast is a region of undulating land that rises from 50-foot (15-metre) hills on the coastal side of the region to 400-foot (120-metre) ones on the western side. The area is between 80 and 100 miles wide and is widest in the southeast. It is covered with sands, from which it takes its name as the white-sands (*zanderij*) region. A small savanna region in the east lies about 60 miles from the coast and is surrounded by the white-sands belt. The sands partly overlie a low crystalline plateau that is generally less than 500 feet in elevation. The plateau forms most of the country's centre and is penetrated by igneous rock intrusions that cause the numerous rapids of Guyana's rivers.

Beyond the crystalline plateau, the Kaieteurian Plateau lies generally below 1,600 feet above sea level; it is the site of the spectacular Kaieteur Falls, noted for their sheer 741-foot initial plunge. The plateau is overlain with sandstones and shales that, in the south, form the extensive Rupununi Savanna region. The Acarai Mountains (Serra Acaraí), which rise to about 2,000 feet, rim the plateau on the southern border, and it is crowned on the western frontier by the Pakaraima Mountains, which rise to 9,094 feet (2,772 metres) in Mount Roraima. The Rupununi Savanna is bisected by the east–west Kanuku Mountains, which rise to almost 3,000 feet.

Major rivers

Drainage. Guyana's four main rivers—the Courantyne, Berbice, Demerara, and Essequibo—all flow from the south and empty into the Atlantic along the eastern section of the coast. Among the tributaries of the Essequibo, the Potaro, Mazaruni, and Cuyuni drain the northwest, and the Rupununi drains the southern savanna. The coast is cut by shorter rivers, including the Pomeroon, Mahaica, Mahaicony, and Abary.

The rivers are part of the watershed of the Amazon and Orinoco rivers, and the headwaters of the Rupununi in Brazil are often confused with those of the Amazon. Drainage is poor, because the average gradient is only one foot per mile, and there are swamps and flooding in the mountains and savannas. The rivers are not suitable for long-distance transportation because they are broken by interior falls, and in the coastal zone their mouths and estuaries are blocked by mud and by sandbars that may occur two to three miles out to sea.

Coastal soils

Soils. The coastal soils are fertile but acidic. The fine-particle, grayish blue clays of the coastal plain are composed of alluvium from the Amazon deposited by the south equatorial ocean current and of much smaller amounts of alluvium from the country's rivers. They overlie white sands and clays and can support intensive agriculture but must be subjected to fallowing to restore fertility. *Pegasse* soil, a type of tropical peat, occurs behind the coastal clays and along the river estuaries, while silts line the banks of the lower rivers. Reef sands occur in bands in the coastal plain, especially near the Courantyne and Essequibo rivers. The rock soils of the interior are leached and infertile, and the white sands are almost pure quartz.

Climate. High temperatures, heavy rainfall with small seasonal differences, high humidity, and high average cloud cover provide climatic characteristics of an equatorial lowland. Temperatures are remarkably uniform. At Georgetown the daily temperature varies from 74° to 86° F (23° to 30° C), and the mean temperature is about 80° F (27° C). The constant heat and high humidity are mitigated near the coast by the trade winds.

Rainfall derives mainly from the movement of the intertropical front, or doldrums. It is heavy everywhere on the plateau and the coast. The annual average at Georgetown is about 90 inches (2,290 millimetres), and on the interior Rupununi Savanna it is about 70 inches. On the coast a long wet season, from April to August, and a short

wet season, from December to early February, are sufficiently well marked on the average, but in the southern savannas the short wet season does not occur. Total annual rainfall is variable, and seasonal drought can occur in July and August when the southeast trade winds parallel the coast. Variations in Guyana's climatic patterns have a determining effect on tropical crop production.

Plant and animal life. Many plants of the coast, such as the mangrove and various saltwater grasses, grow in shallow brackish water and help to protect or extend the land. The wet savanna behind the coast has coarse tufted grasses and a wide scattering of palms, notably the coconut, truli, and manicole. High rain forest, or selva, covers about three-fourths of the land area and is of extraordinary variety and magnificence. Prominent trees include the greenheart and the wallaba on the sandy soils of the northern edge, the giant mora and crabwood on swampy sites, the balata and other latex producers, and many species such as the siruaballi and hubaballi that yield handsome cabinet woods. The interior savanna is mostly open grassland, with much bare rock, many termite hills, and clumps of ita palm.

The rain forest

All forms of animal life are immensely varied and abundant, though few, apart from birds and insects, are normally visible. The tapir is the country's largest land mammal, and the jaguar is the largest and fiercest of the cats, which also include the ocelot; monkeys and deer are the most common animals. Among the more exotic species are the sloth; great anteater; the capybara, or bush pig; and armadillo. Birds include the vulture, kiskadee, blue sacki, hummingbird, kingfisher, and scarlet ibis of the coast and lower rivers; and the macaw, tinamou, bellbird, and cock-of-the-rock in the forest and savanna. The caiman (a reptile similar to the alligator) is the most common of the larger freshwater creatures. The giant anaconda, or water boa, is the largest of the many kinds of

MAP INDEX

Cities and towns

Apoteri . . . . . 04 02 N 58 34 W
Bartica . . . . . 06 24 N 58 37 W
Charity . . . . . 07 24 N 58 36 W
Corriverton . . . . 05 52 N 57 10 W
Enmore . . . . . 06 46 N 57 59 W
Everton . . . . . 06 12 N 57 31 W
Fort Wellington . . 06 24 N 57 36 W
Georgetown . . . . 06 48 N 58 10 W
Isherton . . . . . 02 19 N 59 22 W
Ituni . . . . . 05 30 N 58 14 W
Kamuda Village . . 05 38 N 60 18 W
Karasabai . . . . . 04 02 N 59 32 W
Lethem . . . . . 03 23 N 59 48 W
Linden . . . . . 06 00 N 58 18 W
Mabaruma . . . . . 08 12 N 59 47 W
Mahaicony Village . . . . . 06 36 N 57 48 W
Matthews Ridge . 07 30 N 60 10 W
New Amsterdam . 06 15 N 57 31 W
Orinduik . . . . . 04 42 N 60 01 W
Parika . . . . . 06 52 N 58 25 W
Port Kaituma . . . 07 44 N 59 53 W
Rose Hall . . . . . 06 16 N 57 21 W
Suddie . . . . . 07 07 N 58 29 W
Vreed en Hoop . . 06 48 N 58 11 W

Physical features and points of interest

Abary, *river* . . . . . 06 33 N 57 44 W
Acarai Mountains . . . . . 01 50 N 57 30 W
Amuku Mountains . . . . . 01 45 N 58 20 W
Atlantic Ocean . . 07 30 N 57 00 W
Barama, *river* . . . 07 40 N 59 15 W
Barima, *river* . . . 07 30 N 60 00 W
Berbice, *river* . . . . 06 17 N 57 32 W
Courantyne, *river* . . . . . 05 57 N 57 06 W
Cuyuni, *river* . . . . 06 23 N 58 41 W
Demerara, *river* . . 06 48 N 58 10 W
Ebini, Mount . . . . 05 11 N 59 15 W
Essequibo, *river* . . 06 59 N 58 23 W
Guiana Highlands . . . . . 04 00 N 60 00 W
Ireng, *river* . . . . . 03 33 N 59 51 W
Iwokrama Mountains . . . . . 04 20 N 58 44 W
Kaieteur Falls . . . 05 10 N 59 28 W
Kaituma, *river* . . . 08 11 N 59 41 W
Kalikuri Rapids . . 04 39 N 58 39 W
Kamoa Mountains . . . . . 01 37 N 59 00 W
Kanaima Falls . . . 06 53 N 60 15 W
Kanuku Mountains . . . . . 03 12 N 59 30 W
Kassikaityu, *river* . 01 49 N 58 32 W
Kuyuwini, *river* . . 02 16 N 58 16 W
Kwitaro, *river* . . . . 03 19 N 58 47 W
Mahaica, *river* . . . 06 43 N 57 55 W
Mahaicony, *river* . 06 36 N 57 48 W
Makarapan, Mount . . . . . 04 00 N 58 51 W
Makari, Mount . . . 04 32 N 58 23 W
Mazaruni, *river* . . . 06 25 N 58 38 W
Merume Mountains . . . . . 05 48 N 60 06 W
Moruka, *river* . . . . 07 40 N 58 48 W
New, *river* . . . . . 03 23 N 57 36 W
Ororonoque, *river* . . 02 45 N 57 25 W
Pakaraima Mountains . . . . . 04 05 N 61 30 W
Peaima Falls . . . . 06 21 N 60 36 W
Pomeroon, *river* . . 07 37 N 58 45 W
Potaro, *river* . . . . 05 22 N 58 54 W
Puruni, *river* . . . . 06 00 N 59 12 W
Rewa, *river* . . . . . 03 53 N 58 45 W
Roraima, Mount . . 05 12 N 60 44 W
Rupununi, *river* . . 04 03 N 58 34 W
Rupununi Savanna, *region* . . . . . 03 00 N 59 30 W
Siparuni, *river* . . . 04 50 N 58 49 W
Takutu, *river* . . . . 02 50 N 60 00 W
Tiboku Falls . . . . 05 43 N 59 38 W
Tiger Hill . . . . . 05 39 N 58 23 W
Wairi, *river* . . . . . 08 24 N 59 51 W
Wakenaam Island . . . . . 06 58 N 58 27 W
Wenamu, *river* . . . 06 43 N 61 07 W



snakes, and the bushmaster is the most vicious. Lizards are numerous and include the iguana in the lower rivers. Sharks and stingrays are found offshore. The snapper and grouper are common ocean fish, and shrimps abound in the muddy currents off the coast. The manatee is also common in Guyanaese waters. Among the freshwater fish is the huge pirarucu, which attains lengths up to 14 feet.

Settlement patterns. The country is divided traditionally between the coast, where most of the population is concentrated, and the interior. The coastal population is heterogeneous, its inhabitants descended from the labourers brought in to work the sugarcane plantations. The interior, despite scattered ranching and mining settlements, is largely the province of the indigenous Indians.

Rural society

Guyanese society is predominantly rural, most of the people occupying villages in the coastal region. The highest population concentrations are along the estuary of the Demerara River and between the mouths of the Berbice and Courantyne rivers. Village units are of distinctive rectangular shapes, with the settlement areas nearest the ocean and connected to one another by the coastal highway; each village's farmlands extend inland, often for several miles, and are separated from neighbouring village lands by canals. Villages range in size from several hundred to several thousand persons. The commonly found wood and concrete-block dwellings are usually built on stilts above the flood-prone land and are connected by footbridges to the streets, which are built over the drainage and irrigation canals.

Georgetown is the country's main port and its largest city. Located at the mouth of the Demerara River, it lies below sea level and is protected by dikes along both the river and the sea. Other important towns include the interior bauxite-mining centre of Linden and the market centre of New Amsterdam, located on the mouth of the Berbice River. Agricultural centres, such as the sugarcane plantation of Port Mourant, east of New Amsterdam, and the rice centre of Anna Regina, north of the Essequibo River estuary, provide commercial and marketing functions in the rural areas of the coastal zone.

THE PEOPLE

The indigenous peoples of Guyana are collectively known as Amerindians and constitute about 4 percent of the population. Indian groups include the Warao (Warrau), Arawak, Carib, Wapisiana (Wapishana), Arecuna, the mixed "Spanish Arawak" of the Moruka River, and many more in the forest areas. The Makusí (Macussí or Macushí) are the most prominent of the savanna peoples. Sizable concentrations of Amerindians inhabit the far west along the border with Venezuela and Brazil. They are rarely seen in the populated coastal areas, although a few have interbred with blacks and East Indians. Since 1970, traditional Amerindian lands near the international borders have come under government control, although Amerindians continue to hold village lands informally throughout Guyana's interior.

Chief ethnic groups

The other major elements in the population are predominantly coastal dwellers. Descendants of African slaves form the oldest group; they abandoned the plantations after full emancipation in 1838 to become independent peasantry or town dwellers. The Afro-Guyanese constitute about one-third of the population. The East Indians came mostly as indentured labour from India to replace Africans in plantation work. They form the largest racial group in the country—about half the population—and have been increasing more rapidly than the others. The East Indians are the mainstay of plantation agriculture, and many are independent farmers and landowners, have done well in trade, and are well represented among the professions.

The Chinese and Portuguese also entered originally as agricultural labourers but are now rarely found outside the towns. They are active in business and the professions, and their influence is disproportionate to their numbers; they have not been increasing, however, and together they constitute only a tiny percentage of the population. Europeans other than Portuguese are few, and most are short-term inhabitants. While every kind of racial mixture may be found, mulattoes (persons of mixed white and black ancestry) are by far the most common. Most of them live in towns, and a high proportion are in clerical or professional work.

The major religions are Christian (chiefly Anglican and Roman Catholic) and Hindu. Fundamentalist Protestantism has made inroads in the 20th century, mainly in Georgetown. There is also a sizable minority of Muslims. Animistic religions are still practiced by some of the Amerindian peoples. The official and principal language is English, but a creole patois is spoken throughout the country. Hindi and Urdu are heard occasionally among older East Indians.

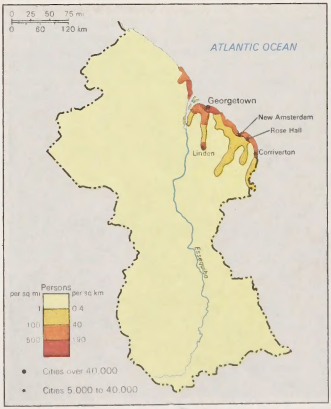
Immigration is no longer important, and by the late 20th century the number of foreign-born, long-term residents was insignificant. Emigration has been a drain on the country's human resources as thousands of persons have left annually, going mainly to the United States, Canada, England, and Caribbean islands. Many of the emigrants have been skilled and professional people whose loss has intensified the country's severe economic problems. East Indians have emigrated in large numbers to flee what they consider political persecution, a number of them having sought part-time work across the Courantyne River in western Suriname.

THE ECONOMY

Since independence Guyana has remained locked into a typical colonial economic dependency on agricultural and mined products, most notably sugarcane and bauxite. Independence brought economic reforms under a socialist-leaning government, but the effect on the old economic cycle has been minimal. Although the government permits a three-sector economy—private, public, and cooperative—the public sector remains heavily dominant.

Government management of the economy has become direct and significant. During the 1970s the government nationalized U.S. and Canadian bauxite holdings; in 1976 it nationalized the vast holdings of the Booker McConnell companies in Guyana, which included coastal sugarcane plantations as well as an array of light manufacturing and commercial enterprises. By the mid-1980s it was estimated that the government controlled directly more than 80 percent of Guyana's economy. All nationalized businesses have been reorganized under the Guyana State Corporation. The state-owned Guyana Sugar Corporation controls the sugarcane plantations, and the Guyana Mining Enterprise Ltd. was established to oversee local mineral production.

Nationalization in the 1970s



Population density of Guyana.

The Guyanese economy has deteriorated under government management policies. Members of the ruling People's National Congress (PNC) political party have been placed in managerial positions, leading to the exodus of former managers and clerical workers. Declining output, a reliance on volatile external commodity markets, and a reduced tax base have all increased financial deficits. External debt has risen precipitously, and a devalued currency has been eroded by speculation in the local black market. Reduced fuel imports have led to widespread power outages, and a government austerity program all but eliminated imported food and consumer goods. Guyana's per capita income (estimated at about $600 in the late 1980s) places it among the world's poorest countries. Improvements in economic conditions became dependent upon foreign aid and a variety of regional and reciprocal trade agreements.

Trade associations

Trade associations have an important influence in Guyanese government. The Trade Union Congress is an association of major unions, among which are the Guyana Mine Workers' Union, which is composed almost exclusively of black workers, and the Guyana Agricultural and General Workers' Union is a predominantly East Indian association.

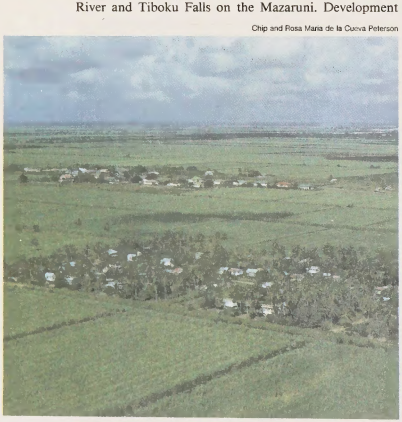
Resources. The most important mineral resource is the extensive bauxite deposits between the Demerara and Berbice rivers. There are also significant deposits of manganese at Matthews Ridge in the northwest, about 30 miles east of the Venezuelan frontier. Diamonds occur in the Mazaruni and other rivers of the Pacaraima Mountains. Gold is found in both alluvial and subsurface deposits. Other minerals include copper, iron ore, molybdenite (the source of molybdenum), nickel, white sand (used in glass manufacture), kaolin (china clay), and graphite. The government has encouraged oil exploration, but no significant reserves have been found.

The main biological resource consists of the hardwoods of the tropical rain forest and especially the greenheart tree, which is resistant to termites, decay, and marine erosion. The shrimps off the coast and a few inland fishes form the basis of the nation's fishing industry, and the grasses of the savanna regions are used for cattle grazing.

Sources of energy

Most of Guyana's energy must be imported; domestic electricity is produced largely by thermal generation and is available only on the coastal plain and along the lower reaches of the rivers. Hydroelectric potential in Guyana is considerable, especially at Tiger Hill on the Demerara River and Tiboku Falls on the Mazaruni. Development is hampered, however, by the remoteness of the falls and the large amounts of capital needed for generation and transmission facilities.

Chip and Rosa Maria de la Cueva Peterson



Sugarcane fields and communities at Ogle, east of Georgetown.

Agriculture, forestry, and fishing. Agriculture is concentrated on the narrow sea-level coastal plain between the Essequibo and Courantyne rivers. Land-use patterns still reflect early Dutch and British water-control techniques. Arable land is laid out in strips between the sea or a river and inland swamps. It is protected on all sides by dikes and canals that are used for both irrigation and drainage. The land reclaimed from the sea is fertile but acidic; lost fertility must be returned to the soil by periodic fallowing or the addition of fertilizers.

Food crops include cassava, corn (maize), bananas, vegetables, and citrus fruits. Cash crops are mainly sugarcane and rice but also include coffee and cacao. Both sugarcane and rice are cultivated through a combination of mechanization and hand labour. Agricultural production increased during the mid-20th century, mainly because mechanization extended cultivable lands, although output stagnated in later decades as the entire economy foundered. East Indian workers overwhelmingly predominate in agriculture.

Livestock production is carried out on the Rupununi Savanna and on the coastal plain. Animals include beef cattle, dairy cattle, pigs, goats, sheep, and poultry.

Forestry activities are hampered by the lack of adequate transportation, the difficulty of cutting the extremely hard wood of Guyana's trees, and the shortage of facilities for the sawing, storing, and shipping of timber. Most of the timber produced for the domestic market and for export is from the greenheart tree.

Many fishing facilities have been improved, and total production has increased as fishing has become a more important part of the economy. Shrimping is carried out primarily for export.

Bauxite production

Industry. Guyana is one of the world's largest producers of bauxite. All alumina (aluminum oxide occurring in hydrated form in bauxite) and most of the bauxite mined is produced at Linden. The rest of the country's bauxite mining takes place on the Berbice River; a processing plant also operates downriver at Everton.

Diamonds continue to be mined by hand and by suction dredges in the interior rivers. Gold is mined by individual prospectors, and large-scale Canadian-financed gold mines were opened late in the 1980s.

The country's many rice mills, like its rice fields, are generally small-scale and individually owned, although there are several large government mills along the coast. Other domestic industries are oriented toward the replacement of consumer imports such as cigarettes and matches, edible oils, margarine, beverages, soap and detergents, and clothing. Refined sugar, stock feeds, and rum and beer are also produced.

Finance and trade. The Bank of Guyana, established in 1965, has the sole right of note issue and acts as banker to the government and other banks. The country's major commercial banks include three local banks and branches of Canadian and Indian banks. Other financial services are provided by the Guyana Cooperative Agricultural and Industrial Development Bank and the New Building Society; insurance companies, most of which are foreign-owned; and more than 1,500 cooperative societies, which serve as savings institutions and offer agricultural credit.

Guyana's major trading partners are the United States, the United Kingdom, and Trinidad and Tobago. Guyana joined the Caribbean Free Trade Association (Carifta) in 1965 and then became a member of the Caribbean Community and Common Market (Caricom), which replaced Carifta in 1973. The major exports are bauxite and alumina, sugar, and rice. Shrimps, diamonds, molasses, rum, and timber are also sold abroad. Major imports include fuels and lubricants, machinery, vehicles, textiles, and foods.

Transportation. The limited road and highway system is partly paved and partly made of burnt clay. The few hundred miles of paved roads are mostly in the coastal zone. The interior has few roads.

Guyana's coastal railway, established in 1848 as South America's first rail line, was discontinued in the 1970s,

ending passenger service. A remaining freight line connects the manganese mines at Matthews Ridge with Port Kaituma on the Kaituma River, and another transports bauxite between Ituni and Linden.

Air and marine transport

Guyana Airways Corporation operates scheduled domestic and international flights. Timehri International Airport, established in 1968 and located 25 miles from Georgetown, is the country's main airport and is served by several international airlines. Domestic commercial and private aircraft, chiefly carrying passengers and equipment, use landing strips and the quieter stretches of rivers.

Barges and small boats carry people and agricultural products in the canals of the coastal estates and villages. Larger boats traverse the estuaries that intersect the coastal plain. A pontoon bridge across the Demerara River opened in 1978; it is the only bridge to link major segments of the coastal plain. Bauxite is loaded into oceangoing ships at Linden and manganese ore at Port Kaituma, but otherwise the country's external trade passes through Georgetown, which maintains connections with the West Indies, Suriname, French Guiana, the United Kingdom, Canada, and the United States.

GOVERNMENT AND SOCIAL CONDITIONS

Government. Guyana became an independent member of the Commonwealth in 1966 and in 1970 became a cooperative republic, involving citizens' organizations in government. Under the constitution of Oct. 6, 1980, executive power is vested in the president, who leads the majority party in the unicameral National Assembly and holds office for the assembly's duration. The president appoints the Cabinet, which is responsible to the National Assembly. The minority members of the assembly elect an opposition leader. The assembly is elected by universal adult suffrage for a term of five years.

The right to vote belongs to all Guyanese citizens 18 years of age or older. Voting is carried out by secret ballot under a system of proportional representation. Votes are cast for lists of candidates compiled by the political parties, and seats are allocated proportionally among the lists.

Political parties

Since independence in 1966, Guyana has been ruled by one party, the People's National Congress. Initially identified with the urban black populace, the PNC essentially established a one-party state under the direction of its first leader, Forbes Burnham. The PNC won power in an election marked by numerous reports of irregularities, many of which were related to the Guyana Defence Force (GDF), a military unit established in 1965 with strong ties to the PNC. Both the GDF and the police force are overwhelmingly black.

The People's Progressive Party (PPP), the PNC's official opposition, is the traditional party of the rural East Indians; smaller parties include the Working People's Alliance (WPA), a newer party founded by the historian Walter Rodney and headed by black labour leaders and intelligentsia allied against alleged PNC corruption.

Local government is administered principally through the Regional Democratic Councils, each led by a chairman; they are elected for terms of up to five years and four months in each of the country's 10 regions.

Guyana has two legal traditions, the British common law and the Roman-Dutch code, the latter now largely relegated to matters of land tenure. The constitution is the supreme law of the land. The court structure consists of magistrate courts for civil claims of small monetary value and minor offenses, the High Court, with original and appellate jurisdiction in civil and criminal matters, and the Court of Appeal, with appellate authority in criminal cases. The Court of Appeal and the High Court together constitute the Supreme Court.

Government control of education

Education. Education is free and compulsory. Primary and secondary instruction are separate, although the lack of facilities makes it necessary to hold some secondary classes in primary schools. In 1976 the government assumed full responsibility for education from nursery school to university. Government authority was extended over all church and private primary schools. Teachers are expected to teach loyalty to both the PNC and socialist objectives. The principal university is the University of Guyana, founded in 1963 and subsequently housed at Turkeyen, in the eastern part of Greater Georgetown. The school has also become politicized, attendance there being contingent upon prospective students completing a year of national service, usually at camps in Guyana's interior. Thus many Guyanese seek education and training abroad. There are also a number of other colleges, including technical and teacher-training schools.

Health and welfare. Health standards declined after independence. Many doctors and other trained personnel have emigrated, and economic austerity programs have reduced supplies of medicine and soap. Food shortages have created widespread malnutrition, especially in Georgetown. Diseases formerly under control, notably beriberi and malaria, had reappeared by the early 1980s, and sanitation problems have also increased.

Under colonial rule public health was centred around government and plantation health clinics. After independence a universal health care system was instituted, and most hospital facilities came under government control. Health problems arise particularly along the easily flooded coast, where the many ditches and ponds provide ideal environments for the spread of disease. A minimal government pension plan for the sick and aged has continued beyond independence, its effectiveness reduced by inflation. Government housing projects, confined mainly to the Georgetown area, have not produced expected results.

The national social structure was inherited from the period of British colonial rule, under which the majority of East Indian and Afro-Guyanese labourers were directed by white planters and government officials. A poorly defined local middle class composed of teachers, professionals, and civil servants, and including a disproportionate number of Chinese and Portuguese, emerged during colonialism. Since independence the PNC political elite has replaced the white plantocracy at the apex of Guyana's social order. The Amerindians remain apart from the country's social structure as they did under the British.

CULTURAL LIFE

Colonial influence

Postindependence Guyanese culture still bears the imprint of its colonial heritage. Guyanese were taught to respect and covet European values during the colonial era, and this has not changed despite government exhortation. Yet ethnic identity continues to be important, with daily life centring around ethnic and family groups; the mother- and grandmother-dominated family among blacks differs from the father-oriented East Indian family. Men of both groups often commute long distances to work along the coastal highway. Daily dress normally does not distinguish one group from another.

Amerindian culture, which remains uninfluenced by national politics, is recognized as an important element in Guyanese museum displays and as an inspiration in local music and painting. Cultural institutions are concentrated in Georgetown, including the Guyana Museum, which includes the Guyana Zoo, with its impressive collection of animals from northern South America. Guyanese writers have made noteworthy contributions to literature; the works of Wilson Harris, A.J. Seymour, and Walter Rodney are among the foremost.

Much recreational activity is based upon the festivities that accompany Hindu, Muslim, and Christian holidays. The Guyanese share the passion for cricket that is prevalent throughout the English-speaking Caribbean.

Media

The government has taken nearly complete control over local news media, including the one radio station and the single daily newspaper. Objections against censorship have been on the rise from opposing political and church groups. In 1988 Guyana's first television station was established under government control.

For statistical data on the land and people of Guyana, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

The first human inhabitants of Guyana probably came into the highlands during the first millennium before

Christ. The Warrau Indians may have arrived first, followed by Arawak and Carib tribes. The early communities practiced shifting agriculture supplemented by hunting. Christopher Columbus sighted the Guyana coast in 1498, and Spain subsequently claimed, but largely avoided, the area between the Orinoco and Amazon deltas, a region long known as the Wild Coast. It was the Dutch who finally began European settlement, establishing trading posts upriver in about 1580. By the mid-17th century they had begun importing slaves from West Africa to cultivate sugarcane. In the 18th century the Dutch, joined by other Europeans, were moving their estates downriver toward the fertile soils of the estuaries and coastal mud flats. Laurens Storm van's Gravesande, governor of Essequibo from 1742 to 1772, coordinated these development efforts.

Formation of British Guiana

Guyana changed hands with bewildering frequency during the wars (mostly between the British and the French) from 1780 to 1815. During a brief French occupation, Longchamps, later called Georgetown, was established at the mouth of the Demerara; the Dutch renamed it Stabroek and continued to develop it. The British took over in 1796 and remained in possession, except for short intervals, until 1814, when they purchased Demerara, Berbice, and Essequibo, which in 1831 were united as the colony of British Guiana.

The slave trade was abolished in 1807, when there were about 100,000 slaves in Berbice, Demerara, and Essequibo. After full emancipation in 1838, black freedmen left the plantations to establish their own settlements along the coastal plain. The planters then imported labour from several sources, the most successful group being indentured workers from India. Indentured labourers who had earned their freedom settled in their own coastal villages near the estates, a process that became established in the late 19th century during a serious economic depression caused by competition with European sugar beet production.

Settlement proceeded slowly, but gold was discovered in 1879, and a boom in the 1890s helped the colony. The North Western District was organized in 1889 and was the cause of a dispute in 1895 when the United States supported Venezuela's claims to the territory. Venezuela revived its claims on British Guiana in 1962, an issue that went to the United Nations for mediation in the early 1980s but which remains unresolved.

The British inherited from the Dutch a complicated constitutional structure. Changes in 1891 led to progressively greater power being held by locally elected officials, but reforms in 1928 invested all power in the governor and the Colonial Office. In 1953 a new constitution—with universal adult suffrage, a bicameral elected legislature, and a ministerial system—was introduced.

From 1953 to 1966 the political history of the colony was stormy. The first elected government, formed by the People's Progressive Party led by Cheddi Jagan, seemed so procommunist that the British suspended the constitution in October 1953 and dispatched troops. The constitution was not restored until 1957. The PPP split along racial lines, Jagan leading a predominately East Indian party and Forbes Burnham leading a party of African descendants, the People's National Congress. In the elections of 1957 and 1961, the PPP was returned with working majorities. From 1961 to 1964 severe rioting involving bloodshed between rival blacks and East Indians and a long general strike led to the return of British troops.

Independence

To answer the PNC allegation that the existing electoral system unduly favoured the East Indian community, the British government introduced for the elections of December 1964 a new system of proportional representation. Thereafter the PNC and a smaller, more conservative party formed a coalition government, led by Burnham, which took the colony into independence under its new name, Guyana, on May 26, 1966. The PNC gained full power in the general election of 1968, which was characterized by questionable rolls of overseas voters and widespread claims of electoral impropriety. On Feb. 23, 1970, Guyana was proclaimed a cooperative republic within the Commonwealth. A president was elected by the National Assembly, but Burnham retained executive power as prime minister. Burnham declared his government to be socialist and in the later 1970s sought to reorder the government in his favour. In 1978 one of the most bizarre incidents in modern history occurred in Guyana when some 900 members of a religious cult in a commune known as Jonestown committed mass suicide at the behest of their leader, Reverend Jim Jones.

In 1980, under a new constitution, Burnham became executive president, with still wider powers, after an election in which international observers detected widespread fraud. In the following years Burnham was faced with an economy shattered by the depressed demand for bauxite and sugar and a restive populace suffering from severe commodity shortages and a near breakdown of essential public services. Burnham enforced austerity measures, and he began leaning toward Soviet-bloc countries for support. Burnham died in 1985 and was succeeded by the prime minister, Hugh Desmond Hoyte, who pledged to continue Burnham's policies. In elections held that year Hoyte won the presidency by a wide margin, but once again charges of vote fraud were raised. In the late 1980s the administration, facing worsening financial and economic problems, moved to liberalize the economy.

For later developments in the history of Guyana, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 964, 966, and 974.

BIBLIOGRAPHY

Physical and human geography: General information can be found in WILLIAM B. MITCHELL *et al., Area Handbook for Guyana* (1969), dated but still useful; K.F.S. KING, *Land and People in Guyana* (1968); and FRANCIS CHAMBERS (comp.), *Guyana* (1989), an annotated bibliography. For statistical information, see the *Statistical Digest* (annual). For natural resources, see VINCENT ROTH (comp.), *Handbook of Natural Resources of British Guiana* (1946), and *Notes and Observations on Animal Life in British Guiana, 1907–1941: A Popular Guide to Colonial Mammalia* (1941); and D.B. FANSHAWE, *The Vegetation of British Guiana, a Preliminary Review* (1952). An early botanical study is WALTER E. ROTH (trans. and ed.), *Richard Schomburgk's Travels in British Guiana, 1840–1844,* 2 vol. (1922–23; originally published in German, 1847–48). Studies of the indigenous population include COLIN HENFREY, *A Gentle People: A Journey Among the Indian Tribes of Guiana* (1964; U.S. title, *Through Indian Eyes,* 1965), which also provides a lively travel account; MARY NOEL MENEZES, *British Policy Towards the Amerindians in British Guiana, 1804–1873* (1977); and ANDREW SANDERS, *The Powerless People: An Analysis of the Amerindians of the Corentyne River* (1987). DWARKA NATH, *A History of Indians in Guyana,* 2nd rev. ed. (1970), examines the East Indian population. RAYMOND T. SMITH, *British Guiana* (1962, reprinted 1980), is an outstanding sociological survey, and *The Negro Family in British Guiana* (1962, reissued 1971), is an anthropological classic. Economic aspects of the sugar industry are dealt with in ALAN H. ADAMSON, *Sugar Without Slaves: The Political Economy of Guyana* (1972), on the 19th century; JAY R. MANDLE, *The Plantation Economy: Population and Economic Change in Guyana, 1838–1960* (1973); WALTER RODNEY, *A History of the Guyanese Working People, 1881–1905* (1981); and CLIVE Y. THOMAS, *Plantations, Peasants, and State: A Study of the Mode of Sugar Production in Guyana* (1984). Views of the country's political situation are presented in LEO A. DESPRES, *Cultural Pluralism and Nationalist Politics in British Guiana* (1967); JACQUELINE ANNE BRAVEBOY-WAGNER, *The Venezuela-Guyana Border Dispute: Britain's Colonial Legacy in Latin America* (1984); HENRY B. JEFFREY and COLIN BABER, *Guyana: Politics, Economics, and Society: Beyond the Burnham Era* (1986); and CHAITRAM SINGH, *Guyana: Politics in a Plantation Society* (1988), a survey of postindependence politics.

History: An early history of Guyana is C.A. HARRIS and J.A.J. DE VILLIERS (comps.), *Storm van's Gravesande: The Rise of British Guiana,* trans. from Dutch, 2 vol. (1911, reprinted 1967), extracts from his dispatches written between 1738 and 1772. ALLAN YOUNG, *The Approaches to Local Self-Government in British Guiana* (1958), deals mainly with the 19th century. BRIAN L. MOORE, *Race, Power, and Social Segmentation in Colonial Society: Guyana After Slavery, 1838–1891* (1987), is a history of race relations. THOMAS J. SPINNER, JR., *A Political and Social History of Guyana, 1945–1983* (1984), provides an overview of recent events. CHEDDI JAGAN, *The West on Trial: The Fight for Guyana's Freedom,* rev. ed. (1972, reissued 1980), is a vivid account of preindependence turmoil by a former prime minister. LATIN AMERICAN BUREAU, *Guyana: Fraudulent Revolution* (1984), takes a closer look at Burnham's government.

(B.C.Ri./Ed.)

Gymnosperms

Gymnosperms are a group of vascular plants whose seeds are not enclosed by a ripened ovary (fruit). In 1825 the Scottish botanist Robert Brown distinguished gymnosperms from the other major group of seed plants, the angiosperms, whose seeds are surrounded by an ovary wall. The seeds of many gymnosperms (literally, "naked seed") are borne in cones and are not visible. These cones, however, are not the same as fruits. During pollination, the immature male gametes, or pollen grains, sift among the cone scales and land directly on the ovules (which contain the immature female gametes) rather than on elements of a flower (the stigma and carpel) as in angiosperms. Furthermore, at maturity, the cone expands to reveal the naked seeds. Gymnosperms were considered at one time to be a class of seed plants, called Gymnospermae, but taxonomists now tend to recognize four distinct divisions of extant gymnospermous plants (Coniferophyta, Cycadophyta, Ginkgophyta, Gnetophyta) and to use the term gymnosperms only when referring to the naked-seed habit. Some of the divisions of gymnosperms are not closely related to others, having been distinct groups for hundreds of millions of years. Currently, about 60–70 genera are recognized, with a total of 700–800 species. Gymnosperms are distributed throughout the world, with extensive latitudinal and longitudinal ranges.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 313, and the *Index*.

This article is divided into the following sections:

The gymnosperms: an overview 447
- General features 447
  - Diversity in size and structure
  - Distribution and abundance
  - Importance to humans and ecology
- Natural history 448
- Form and function 449
  - General features
  - Stems
  - Leaves
  - Roots
  - Strobili
- Evolution and paleobotany 450
- Classification 452
  - Distinguishing taxonomic features
  - Annotated classification
  - Critical appraisal

The gymnosperm divisions 452
- Coniferophytes 452
  - General features
  - Natural history
  - Form and function
  - Classification
- Cycadophytes 457
  - General features
  - Natural history
  - Form and function
  - Classification
- Gnetophytes 461
  - General features
  - Form and function
  - Classification
- Ginkgophytes 464
  - General features
  - Form and function
  - Classification

Bibliography 465

THE GYMNOSPERMS: AN OVERVIEW

GENERAL FEATURES

Diversity in size and structure. Among the gymnosperms are plants with stems that may barely project above the ground and others that develop into the largest of trees. Cycads resemble palm trees, with fleshy stems and leathery, featherlike leaves. The tallest cycads reach 19 metres (62 feet). *Zamia pygmaea*, a cycad native to Cuba, has a trunk less than 10 centimetres (four inches) in height. Of the gnetophytes, *Ephedra* (joint fir) is a shrub and some species of *Gnetum* are vines, while the unusual *Welwitschia* has a massive, squat stem that rises a short distance above the ground. The apex is about 60 centimetres in diameter. From the edge of the disk-shaped stem apex arise two leathery, straplike leaves that grow from the base and survive for the life of the plant. Most gymnosperms, however, are trees. Of the conifers, the redwoods (*Sequoia*) exceed 100 metres in height and, while *Sequoiadendron* (giant redwood) is not as tall, the trunk is more massive.

Distribution and abundance. Although since the Cretaceous period (144 to 66.4 million years ago) gymnosperms have been gradually displaced by the more recently evolved angiosperms, they are still successful in many parts of the world and occupy large areas of the Earth's surface. Conifer forests, for example, cover vast regions of northern temperate lands in North America and Eurasia. In fact, they grow in more northerly latitudes than do angiosperms. Vascular plants that occur at the highest altitudes are the gnetophyte *Ephedra*. Land in the Southern Hemisphere is rich in conifer forests, which tend to be more abundant at higher altitudes. Gymnosperms that occupy areas of the world with severe climatic conditions are adapted to conserving water; leaves are covered with a heavy, waxy cuticle, and pores (stomata) are sunken below the leaf surface to decrease the rate of evaporation.

Cycads are distributed throughout the world but are concentrated in equatorial regions. As a natural population, *Ginkgo* originally appeared to have been confined to mountains of southeastern China; extensive artificial propagation has altered this natural distribution. Distribution of gymnosperms in the distant past was much more extensive than at present. In fact, gymnosperms were dominant in the Mesozoic era (245 to 66.4 million years ago), during which time some of the modern families originated (Pinaceae, Araucariaceae, Taxodiaceae).

Importance to humans and ecology. Some of the oldest living things on earth are gymnosperms. Redwoods live for thousands of years, and some specimens of the bristlecone pine, found in the White Mountains of California, approach 5,000 years in age.

Gymnospermous plants are widely used as ornamentals. Conifers are often featured in formal gardens and are used for bonsai. Yews and junipers are often low-growing plants cultivated for ground cover. Conifers are effective windbreaks, especially those that are evergreen. Cycads are used as garden plants in warmer latitudes, and some may even thrive indoors. Their leathery green foliage and sometimes colourful cones are striking. *Ginkgo* is a hardy tree, and although it once approached extinction, it is now cultivated extensively and survives such challenging habitats as the streets of New York City. Some gymnosperms are weedy in that they invade disturbed areas or abandoned agricultural land. Pines and junipers are notorious invaders, making the land unusable.

Horticultural uses

Most of the commercial lumber in the Northern Hemisphere is derived from the trunks of conifers such as pine, Douglas fir, spruce, fir, and hemlock. *Araucaria*, kauri, and *Podocarpus* are important conifers of the Southern Hemisphere used for lumber. The wood is straight-grained, light for its strength, and easily worked. Wood of gymnosperms is often called softwood to differentiate it from the hardwood angiosperms. Wood of angiosperms typically has more kinds of elements than does softwood of gymnosperms. In addition to its use in building construction, gymnospermous wood is used for utility poles and railroad ties. Aromatic wood of cedar is frequently used in the construction of closets or clothes chests and apparently repels cloth-eating moths. Most plywood is gymnospermous. Fibres of conifers make up paper pulp and may occasionally be used for creating artificial silk or other textiles. Conifers are frequently planted in reforestation projects. Conifer bark is often the source of compounds involved in the leather tanning industry. Bark is also used extensively as garden mulch.

Resins

From conifer resins are derived turpentine and rosin. A hardened form of resin from a kauri (*Agathis australis*), called copal, is used in the manufacture of paints and varnishes. Some resins, such as balsam (from hemlock) and dammar (from *Agathis*) are used in the preparation of mounting media for microscope slides. Resins may also have medicinal uses. Many types of amber are derived from fossilized resin of conifers. Commercially useful oils are derived from such conifers as junipers, pines, hemlock, fir, spruces, and aborvitae. These oils serve as air fresheners, disinfectants, and scents in soaps and cosmetics.

Seeds are often food sources. Pine seeds are a delicacy eaten plain or used as a garnish on bakery products. Seeds of *Ginkgo* and cycads may be poisonous unless detoxified. "Berries" (in reality the fleshy cones) of juniper are used to flavour gin.

NATURAL HISTORY

In all living gymnosperm groups, the visible part of the plant body, *i.e.*, the growing stem and branches, represents

From (A,B,E) *The Plant Kingdom* by William H. Brown, © copyright 1935, by William H. Brown, published by Ginn and Company, used by permission of Silver, Burdett & Ginn Inc.; (C,F,J,L,R) T.E. Weier, *Botany: An Introduction to Plant Biology*, copyright © 1982 by John Wiley & Sons, Inc.; (C,H,J,K) redrawn after J.J. Coulter and C.J. Chamberlain, *Morphology of the Gymnosperms*, © 1917 by the University of Chicago; (D) R.E. Torrey, *General Botany for Colleges*, © 1932, renewed 1960, fig. 241A, reprinted by permission of Prentice Hall, Inc., Englewood Cliffs, N.J.

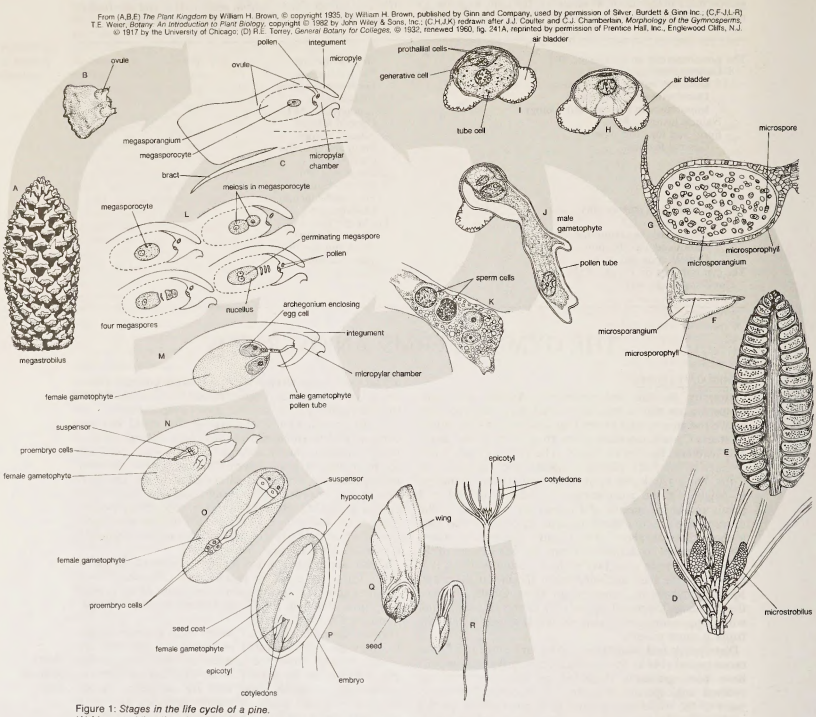


Figure 1: *Stages in the life cycle of a pine.*
(A) Megastrobilus (female cone) at the time of pollination. (B) Ovuliferous scale showing two ovules. (C) Cross section of an ovuliferous scale showing the megasporocyte and the pollen in the micropylar chamber. (D) Microstrobilus. (E) Cross section of a microstrobilus. (F) Scale of a microstrobilus. (G) Cross section of a scale. (H) Winged pollen grain. (I) Development of a male gametophyte in a pollen grain. (J) Pollen tube. (K) Tip of the pollen tube. (L) Stages in megaspore production. (M) Fertilization, showing a female gametophyte with an egg cell and pollen tube. (N) Embryo development. (O) Proembryo. (P) Section of a mature seed showing the seed coats, female gametophyte, and embryo sporophyte. (Q) Winged seed. (R) Germinating seedlings.

Alternation of generations

the sporophyte, or asexual, generation (Figure 1), rather than the gametophyte, or sexual, generation.

In most gymnosperms the pollen cones, called microstrobili, contain reduced leaves called microsporophylls. Microsporangia, or pollen sacs, are borne on the lower (abaxial) surfaces of the microsporophylls. The number of microsporangia may vary from two in many conifers to hundreds in some cycads. Within the microsporangia are cells, called microsporocytes, which undergo meiotic division to produce haploid microspores.

The gametophyte phase begins when the microspore, while still within the microsporangium, begins to germinate to form the male gametophyte. A single microspore nucleus divides by mitosis to produce a few cells. At this stage the male gametophyte (called a pollen grain) is shed and transported by wind or insects.

Ovulate cones, called megastrobili, may be borne on the same plant that bears microstrobili (as in conifers) or on separate plants (as in cycads and *Ginkgo*). A megastrobilus contains many ovuliferous scales, called megasporophylls, that contain megasporangia, within which a single cell (a megasporocyte) undergoes meiotic division to produce four haploid megaspores. Typically three degenerate, leaving one functional megaspore, which is retained within the megasporangium. The female gametophyte begins development within the megaspore. Initial divisions of the megaspore nucleus are mitotic without accompanying divisions of the cytoplasm. As the number of free nuclei multiplies, the megasporangium, integument, and megaspore wall expand. Cell walls eventually develop around the nuclei. At this stage the ovule is ready to be fertilized.

Pollination

Before fertilization can take place, however, the mature male gametophyte (the pollen grain) must be transported to the female gametophyte—the process of pollination. In many gymnosperms, a sticky "pollination droplet" oozes from a tiny hole in the megasporangium (the micropyle) and pollen grains are caught in the droplet. The droplet is then resorbed through the micropyle into the megasporangium. The pollen grain settles on the surface of the megasporangium, where the male gametophyte further develops. A pollen tube emerges from the grain and grows through the megasporangium toward the multicellular egg-containing structure called the archegonium. The egg and sperm continue to mature, the nucleus of the latter undergoing additional divisions resulting in two male gametes, or sperm. (Sperm cells have flagella in cycads and *Ginkgo* but not in any other seed plants.) By the time the pollen tube reaches the archegonium, both the egg and sperm are fully mature, and the egg is ready to be fertilized.

The nuclei of the two sperm are injected into the egg cell; one nucleus dies, and the other unites with the egg nucleus to form a diploid zygote. The nucleus of the fertilized egg begins the development of a new sporophyte generation by dividing a number of times; the resulting multicellular structure becomes the embryo of the seed (Figure 2). Food for the developing embryo is provided by the massive, starch-filled female gametophyte that surrounds it. The time interval between pollination and maturation of the embryo into a new sporophyte generation varies among different groups, ranging from a few months to over one year (in pine, for example). The integument develops into the seed coat, while the female gametophyte is a source of food for the developing embryo during germination.

In some gymnosperms (*e.g.*, cycads, *Ginkgo*) the seed coat (sarcotesta) consists of two layers. In some cycads the sarcotesta is brightly coloured. The sarcotesta of *Ginkgo* seeds is foul-smelling when ripe. Attached to the seed coat in pine and related conifers is a thin membranous wing-like structure, which remains with the seed at its release and serves as a wing that may assist in the distribution of the seed. Members of the order Taxales have a fleshy structure, an aril, surrounding the actual seed. Cones of juniper are fleshy, and the entire fleshy unit drops off or is picked off by birds. Juniper seeds pass through the digestive tracts of birds and are thus distributed effectively.

Maturation of the embryo

At maturity, a gymnosperm embryo has two or more seed leaves (the cotyledons). Cycads, *Ginkgo*, and gnetophytes have two cotyledons in the embryo; pine and other conifers may have several (eight is common; some have as many as 18). Below the attachment point of the cotyledons is the hypocotyl, which emerges through the seed coat during germination, bends downward, and eventually establishes the root system. Above the attachment point of the cotyledons is the epicotyl, the tip of which contains the shoot tip and leaves. In cycads and *Ginkgo* the cotyledons remain within the seed and serve to digest the food in the female gametophyte and absorb it into the developing embryo. Conifer cotyledons typically emerge from the seed and become photosynthetic after digesting and absorbing the food in the female gametophyte.

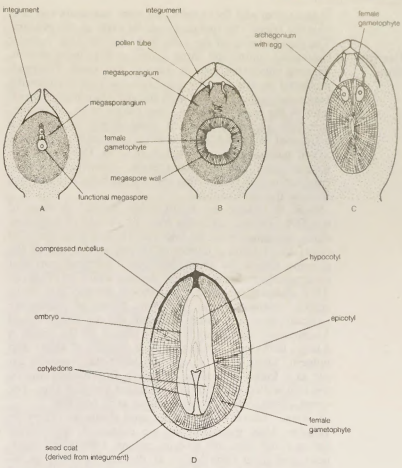


Figure 2: *Stages in the development of the seed in gymnosperms.*
(A) Young ovule at the time of megaspore formation. (B) Development of the megaspore and germination of pollen tubes. (C) Mature ovule at the time of fertilization. (D) Mature seed.

Fig. 30–1 from *Botany* by Peter M. Ray, Taylor A. Steeves, and Sara A. Fultz, copyright © 1983 by Saunders College Publishing, a division of Holt, Rinehart and Winston, Inc. reprinted by permission of the publisher

FORM AND FUNCTION

General features. The visible part of the gymnospermous plant body represents the sporophyte generation. Typically, a sporophyte has a stem with roots and leaves and bears the reproductive structures. The vascular system contains two conducting tissues, the xylem and phloem. The xylem is a tissue containing nonliving cells whose walls form a conducting system of "pipes" through which water and minerals are conducted from the roots to the shoots. The sturdy nature of the xylem makes it useful in support as well. The phloem, like the xylem, is a conducting tissue; its cells, however, are living and distribute the sugars, amino acids, and organic nutrients manufactured in the leaves to the nonphotosynthetic tissues of the plant. When the plant is actively growing, the phloem may also conduct stored nutrients from the roots to the developing shoots.

Xylem and phloem

Stems. The stems, roots, and branches of vascular plants undergo secondary growth, which takes place from stem and branch growth tissue, called the vascular cambium. Stems of conifers are characteristically woody with a dense mass of secondary xylem. They are usually branched, with basal branches dropping off as the stem elongates, resulting in a main stem that is often tall and straight. The wood is simpler than that of angiosperms; it consists primarily of elongated water- and mineral-conducting cells (tracheids)

in the xylem and living cells that store materials and provide for lateral conduction (vascular rays) in the phloem. The growth tissue of the stem and branches (the vascular cambium) contributes more xylem each growing season, forming concentric growth rings in the wood. Tracheids produced by the vascular cambium early in the growing season are larger, and the walls thinner, than those formed later in the growing season. This results in the characteristic light and dark bands of wood. Some conifers have additional cell types, such as fibres and axially elongated xylem parenchyma cells that store food. Phloem is also simpler than that of angiosperms, consisting of food-conducting cells (sieve cells) and storage cells. Phloem rays traverse the phloem tissue.

Growth rings

Stems of *Ginkgo* are anatomically similar to those of conifers. *Ginkgo* and cedar have two kinds of branches: elongated major branches and dwarf lateral branches. The dwarf shoot bears a cluster of leaves; at the end of the growing season the shoot develops a terminal bud that elongates the following year to produce a new set of leaves. After several years these dwarf shoots develop into short, stubby outgrowths from the stem. Stems of cycads are typically short and squat, although the Australian cycad *Macrozamia hopei* may reach 19 metres. In the centre is a large, fleshy pith surrounded by a cylinder of xylem and phloem. There never is as much secondary vascular tissue as is found in conifers, however. Interspersed among the thin-walled tracheids are abundant vascular rays. The wood, consequently, is not as dense as in conifers.

Leaves. Leaves of gymnospermous plants are extremely variable. Most gymnosperms are evergreen, with leaves lasting more than one growing season. Others are deciduous and drop their leaves at the end of every growing season. Bald cypress (*Taxodium*), larch (*Larix*), and dawn redwood (*Metasequoia*) are examples of deciduous conifers. *Ginkgo* also sheds its leaves in the autumn. Among the conifers, leaves are always simple; that is, the blade is a single unit. Leaves may be small and scalelike (*e.g., Thuja*) or needlelike (*Abies, Picea, Pinus*) or have a broad blade (*Araucaria, Agathis*). In some conifers (*Taxodium*) small branch fragments with numerous needlelike leaves are dropped at the end of the growing season.

Cycad leaves are compound, with thick, leathery leaflets borne in a featherlike (pinnate) arrangement on a main axis (rachis). Produced among the normal photosynthetic leaves of cycads are reduced, pointed, stiff, scalelike leaves called cataphylls. These contribute to the persistent "armour" on the trunk surfaces.

Ginkgo resembles an angiospermous tree in that the woody stem is frequently and irregularly branched and bears broad leaves, which are fan-shaped with dichotomously branched veins. The leaves of *Gnetum* look much like those of dicotyledonous angiosperms. *Ephedra* has small, scalelike leaves.

In certain conifers, such as pine and cedar, leaves are borne on dwarf lateral branches that do not elongate, but are telescoped. In cedar, the dwarf lateral shoots grow forward each year producing a new cluster of needles each season. In pine, however, the number of needles per cluster is small (one to eight) and no more needles are produced on the dwarf shoot after the first year.

Ephedra and *Gnetum* do not produce extensive vascular cylinders; *Ephedra* is shrubby, while some species of *Gnetum* are vines. The stem of *Welwitschia* is somewhat turnip-shaped and does not project very high above the ground. The apex is broad and concave, with leaves and reproductive structures borne along the edges. *Gnetum,* unlike most gymnosperms, has vessels in the xylem. Perforations (pores) at the ends of the conducting elements connect them to adjacent elements.

Roots. Filaments of the fungi called endomycorrhizae live within the cells of the roots of certain gymnosperms, especially conifers. Endomycorrhizal fungi are apparently parasitic, but not destructively so. In cycads, blue-green algae grow in nodules in the roots. These roots may grow opposite to the force of gravity and may form corallike masses on the ground surface, hence the term "coralloid roots." It is thought that these fungi and blue-green algae fix atmospheric nitrogen into a form usable by the plant.

Endomycorrhizae

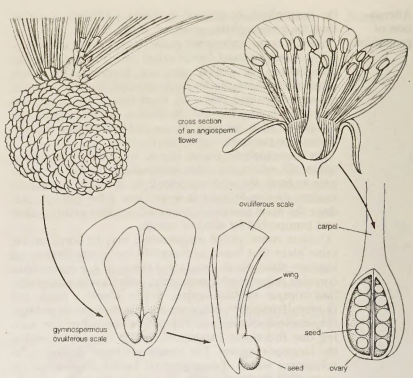


Figure 3: *Comparison of gymnosperm cone and angiosperm flower.*
The seed in gymnosperms is borne exposed (the example above is pine). The angiosperm seed is enclosed within a carpel.
Modified from *Life: An Introduction to Biology* by George Gaylord Simpson and William S. Beck. © 1957, 1965 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission

Strobili. In most conifers the pollen-bearing and ovule-bearing components (the microsporangia and megasporangia, respectively) are borne on the same plant, though separately (monoecious). A pollen-bearing cone, the microstrobilus (Figure 3), consists of a central axis on which are borne, in a close helical arrangement, reduced, fertile leaves (the microsporophylls). On the lower surfaces of the microsporophylls are borne elongated microsporangia; two microsporangia per microsporophyll are common, but some genera have more. The ovulate cone, the megastrobilus, is more complex than the microstrobilus. The megastrobilus bears seeds on flattened dwarf branches, all of the parts of which are fused (ovuliferous scales). Subtending the ovuliferous scale on the cone axis is a reduced scale leaf, or bract. In some conifers the bract is not recognizable because it has been fused to the ovuliferous scale.

In *Ginkgo,* microsporangia and megasporangia are borne on separate trees (*i.e.,* it is dioecious). A *Ginkgo* microstrobilus is borne on a dwarf shoot among the fan-shaped leaves. The microstrobilar axis bears stalked appendages at the ends of each of which are two microsporangia directed downward. A megastrobilus is not recognized as such. Among the leaves of a dwarf shoot on a plant other than one bearing microstrobili are borne elongated, slender stalks, each with a pair of terminal ovules. Usually only one ovule matures into a seed.

Cycads also are dioecious, and all genera bear microstrobili consisting of an axis with microsporophylls inserted in a close, helical arrangement. The microsporophylls are reduced leaves with abaxial sporangia. In the genus *Cycas,* ovules are borne among the edges of the stalk of a reduced leaf with a bladelike region still present. These modified leaves, or megasporophylls, are clustered at the apex of the plant but not arranged in a cone. All other genera of cycads, however, have megastrobili, with the megasporophylls reduced and not leaflike in appearance. Each megasporophyll has a stalk with an expanded distal portion, on the inner face of each of which develop two seeds.

Ephedra, of the Gnetophyta, may have both microstrobili and megastrobili on the same plant, or they may occur on separate plants. The two remaining genera of Gnetophyta, *Gnetum* and *Welwitschia,* are dioecious.

EVOLUTION AND PALEOBOTANY

First seed plants

The first seed plants to have evolved were gymnospermous in the sense that the seeds were naked. The earliest

seedlike bodies are found in rocks of the Late Devonian epoch (374 to 360 million years ago). During the course of the evolution of the seed habit, a number of morphological modifications were necessary. First, all seed plants are heterosporous: two kinds of spores (microspores and megaspores) are produced by the sporophyte. Hence, it is assumed that the ancestors of seed plants must have been heterosporous. Sporangia of plants that do not bear seeds typically lack an integument. The origin of the integument in seed plants was made clear by a study of Early Carboniferous (360 to 320 million years ago) ovules from Scotland. One example, *Genomosperma kidstonii,* consists of an elongated megasporangium with one functional megaspore. Arising from the base of the megasporangium were eight elongated, fingerlike processes that loosely surrounded the megasporangium. In a related species, *G. latens,* these eight fingerlike processes were fused at the base into a cup, with eight free tips. These tips tended to cover the megasporangium rather closely, as opposed to the flared appendages in *G. kidstonii.* Ultimately these fingerlike appendages were almost completely fused into a continuous integument surrounding the megasporangium. A small hole, the micropyle, is left at the apex of the megasporangium where the integument does not quite cover its tip.

In searching for seed-plant ancestors it is necessary to look for a heterosporous type of plant bearing leaves and also having an internal structure similar to that of seed plants. The extinct division Progymnospermophyta provides such an ancestral condition. The best-known progymnosperm is the late Devonian *Archaeopteris,* originally assumed to be a fern with wedge-shaped, subdivided leaflets (pinnules) and sporangia borne on appendages taking the place of pinnules. What was first interpreted as the frond axis was shown to have internal structure like that of *Callixylon,* known as late Devonian stems and wood fragments assumed to be gymnospermous. *Callixylon* wood is like that of many conifers, consisting of tracheids and vascular rays, with closely spaced circular bordered pits on the radial walls of the tracheids. Pits are clustered, separated from other clusters by an area of the wall lacking pits. What were assumed to be pinnae of the frond of *Archaeopteris* are actually branches, and the so-called pinnules are helically arranged leaves. At least some species are known to be heterosporous, hence *Archaeopteris* has many of the features to be anticipated in a seed-plant ancestor.

From progymnosperms such as *Archaeopteris* could have arisen more than one group of gymnosperms. Those with compound leaves (*e.g.,* pteridosperms and cycads) have leaves that would correspond to a flattened branch system of *Archaeopteris.* Those with simple leaves (*e.g.,* conifers) have leaves that are probably the equivalent of the wedge-shaped *Archaeopteris* leaves.

Earliest gymnosperms

The earliest recognized group of gymnospermous seed plants are members of the division Pteridospermophyta (pteridosperms, or seed ferns). These plants originated in the late Devonian and were widespread toward the end of the Paleozoic era (570 to 245 million years ago). Descendants persisted into the Mesozoic. In habit, Paleozoic seed ferns resembled some progymnosperms in that they were small trees with fernlike leaves (the equivalent of a progymnospermous flattened branch) bearing seeds. While Paleozoic seed ferns resembled ferns externally, the internal structure was like that of gymnosperms. Secondary vascular tissues were common in stems of seed ferns. The wood, however, was composed of thin-walled tracheids and abundant vascular rays, suggesting that stems were fleshy like those of cycads. Pteridosperm seeds were very similar to those of cycads. Many were large, with an outer, softer seed coat and a harder, inner seed coat. Within an ovule ready for fertilization was a massive female gametophyte with several archegonia. There has been one report of a pollination droplet in a Carboniferous pteridosperm ovule and a report of a pollen tube emerging from a pollen grain in the micropyle of a seed-fern ovule, suggesting that transport of the sperm through a pollen tube (siphonogamy) was in existence as far back as the Paleozoic. In some pteridosperms the seed was contained within a cupule; some botanists interpret the cupule as a precursor of an angiosperm carpel. Pollen grains, however, landed directly on the micropyle of the ovule. Pollen-bearing organs were variable among the pteridosperms; in many cases the microsporangia were elongated and fingerlike and were produced in clusters or were fused into compound organs. Mesozoic seed ferns are less well defined, and the concept of pteridospermy is used loosely to refer to plants with fernlike foliage bearing seeds.

It is generally conceded that from the pteridosperms arose members of the division Cycadophyta. The first cycads appeared in the Permian period (286 to 245 million years ago). Some of these presumed cycads differ from extant members in that megasporophylls were undivided, unlike those of *Cycas,* considered to be primitive among cycads, in which the distal portion of the megasporophyll may be pinnately divided. Other Permian megasporophylls, from China, are more like those of *Cycas.* Cycad remains, especially leaves, are abundant in Mesozoic rocks. For this reason paleobotanists often refer to the Mesozoic era as the "age of cycads." The earliest well-known cycads appear to have had slender stems, sometimes branched, with leaves not borne close together, unlike the situation in extant cycads in which leaves are densely crowded at the apex of the plant. There is evidence that these earliest cycads were deciduous. Megasporophylls of Mesozoic cycads are essentially like those of extant cycads. The megasporophyll of the Triassic *Palaeocycas* is like that of *Cycas.* Jurassic megasporophylls are like those of most other cycads. Extant cycads are now limited in geographic distribution to the warmer parts of the earth.

First cycads

Coexisting with the cycads during the Mesozoic was another group of gymnosperms, the cycadeoids (division Cycadeoidophyta—sometimes called the Bennettitophyta). Although they are superficially similar in habit to the cycads, with a squat trunk and often pinnately divided leaves, their reproductive structures were different, and their actual relationship is not close. Typically seeds were borne on the surface of a fleshy receptacle. Among the seeds were sterile structures, called interseminal scales, that held the seeds tightly together. Pollen organs were quite similar among the forms in the sense that all had a whorl of modified leaves (microsporophylls) on which were borne compound microsporangia.

Conifers (division Coniferophyta) appeared first toward the end of the Late Carboniferous epoch (320 to 286 million years ago). Some of the earliest conifers (class Cordaitopsida) were trees with long, strap-shaped leaves. Trunks were similar to those of extant conifers, with dense, compact wood; small, thick-walled tracheids; and narrow vascular rays. Reproductive axes were slender, bearing narrow bracts in the axils of which were small, budlike shoots with helically arranged scales. On some of the topmost scales were borne elongated microsporangia. Buds on other axes bore ovules instead of microsporangia.

Earliest conifers

By the late Paleozoic there came into existence another group of extinct conifers, the Voltziales (class Coniferopsida). In general habit they must have resembled some of the extant araucarias (*e.g.,* Norfolk Island pine), with whorled, flattened branches bearing helically arranged, needlelike leaves. Reproductive axes were generally homologous with those of the Cordaitales, but they were more compact, with the bracts on the ovule-bearing axes obscuring the axillary fertile buds. During the end of the Paleozoic and in the early Mesozoic, these axillary buds underwent further transformation. The sterile, non-seed-bearing part became flattened, with the scales fused together. The ovule-bearing portion was situated toward the upper surface (away from the bract). The ovuliferous scale of a conifer seed cone, then, may be interpreted as an axis bearing bracts in the axils of which are modified woody ovuliferous scales derived from lateral buds.

Modern families of conifers began to appear in the Mesozoic era. Members of the Taxodiaceae, the family to which redwoods and bald cypress are assigned, appeared first in the Jurassic period. *Metasequoia,* the dawn redwood, is also a member of this family. Discovered first as fossils in Miocene (23.7 to 5.3 million years ago) deposits, it was assumed to have become extinct until it was discovered growing in Szechwan province in China. Its distribution

in the late Mesozoic and Tertiary (66.4 to 1.6 million years ago) was throughout the Northern Hemisphere. The plant has since been introduced to a variety of places in the world.

During late Triassic times there existed a type of conifer (*Compsostrobus*) that had many features of the Pinaceae. Seed cones had woody ovuliferous scales subtended by bracts with two ovules on the upper surface of each ovuliferous scale. More typical pinaceous remains occurred later in the Mesozoic. Coniferophytes were the dominant vegetation just before the appearance of the angiosperms.

Ginkgo

The division Ginkgophyta, represented today by only one living species, *Ginkgo biloba,* was much more widespread in past ages. Gymnosperms that were presumed to be ginkgophytes existed as far back as the Permian period. In Mesozoic rocks, *Ginkgo* leaves are commonly found throughout the world. The oldest fossil ginkgophytes had leaves much more dissected than the typical *Ginkgo* leaf, resembling more closely the leaves found on new growth in extant ginkgoes.

The fossil record of the division Gnetophyta is obscure, and its origin is not clear. Pollen grains similar to those of *Ephedra* and *Welwitschia* are found as far back as the Permian period. Megafossil remains of possible gnetophytan plants occur in Late Cretaceous (97.5 to 66.4 million years ago) deposits. The plant is unlike any existing one, but venation of the foliage is similar to that of leaves of *Welwitschia.* Pollen grains are typically gnetophytan.

CLASSIFICATION

Distinguishing taxonomic features. Gymnosperms differ from angiosperms most obviously on the basis of the naked-seed habit in the former and the enclosure of seeds within a fruit in the latter. The pollen grain of gymnosperms, when shed from the microsporangium, has more than two cells (three in cycads and four in *Ginkgo* and conifers). Furthermore, gymnosperm pollen lands on the ovule directly, whereas in angiosperms pollen lands on the stigma of a carpel and germinates there, with the pollen tube growing through stigmatic and stylar tissues to reach the ovule. In angiosperm pollen tubes, a total of three cells make up the male gametophyte; gymnosperms have more. The female gametophyte in gymnosperms is much larger than that of angiosperms and serves as the source of food for the developing embryo sporophyte. The female gametophyte of angiosperms consists normally of just a few cells. Both sperm cells in an angiosperm pollen tube are functional, one fertilizing the egg, the other joining with two other nuclei of the female gametophyte. Division of this latter cell forms a multicellular tissue (endosperm) in which food is stored for the embryo. Gymnospermous ovules typically have only one integument; most angiosperms have two.

Annotated classification. Older classifications considered all seed plants to be assignable to a single division, Spermatophyta. The Angiospermae and Gymnospermae were two classes that made up the division. More recent classifications recognize that the characteristic of naked seeds is not important enough to be used to tie all plants with that feature into one group. Classification of gymnosperms now recognizes four separate divisions. Groups marked with a dagger (†) are known only from fossils and have no living members.

†DIVISION PTERIDOSPERMOPHYTA
Late Devonian to Jurassic; seed plants resembling tree ferns with compound, frondlike leaves; seeds and microsporangia borne on the leaves; most stems with secondary vascular tissues.

DIVISION CYCADOPHYTA
Permian to the present; palmlike plants; leaves usually pinnately compound; dioecious; seeds borne in megastrobili with reduced megasporophylls, each bearing inwardly directed seeds (except for the living genus *Cycas*); microstrobili with microsporophylls bearing abaxial microsporangia; 11 extant genera.

†DIVISION CYCADEOIDOPHYTA (BENNETTITOPHYTA)
Triassic (Permian?) to the Cretaceous; cycadlike plants; leaves usually pinnately compound (some entire); ovules borne on the surface of a fleshy receptacle and separated by interseminal scales; microsporangia compound and borne on fingerlike structures fused at the base.

†DIVISION GLOSSOPTERIDOPHYTA
Permian; trees with tongue-shaped leaves with net venation; trunks with compact conifer-like wood; seeds borne on, or associated with, leaves, microsporangia borne on tongue-shaped leaves.

DIVISION CONIFEROPHYTA
Late Carboniferous to the present; woody plants, usually trees, with simple leaves; wood compact; microstrobilus bearing microsporophylls with elongated abaxial microsporangia; seeds borne on megastrobili; ovule with a single integument.

†Class Cordaitopsida
Mississippian to the Permian (or perhaps into the Mesozoic); trees; leaves elongated, strap-shaped; wood compact, coniferlike; fertile shoots slender and elongated with fertile buds borne in the axils of reduced leaves or bracts.

Class Coniferopsida
Late Carboniferous to the present; mostly trees; leaves scalelike, needlelike, or flat and bladelike; wood compact; microstrobili with reduced microsporophylls with abaxial sporangia; megastrobili usually bearing woody ovuliferous scales derived from flattened dwarf branches; seeds borne on the upper surface; 50 genera.

Class Taxopsida
Triassic to the present; trees or shrubs; leaves needlelike; microstrobili with microsporophylls bearing abaxial microsporangia; seeds not in megastrobili but terminating dwarf shoots and surrounded by a fleshy aril; 5 genera.

DIVISION GINKGOPHYTA
Permian to the present; dioecious trees with fan-shaped leaves; bilobed or with more lobes, especially in fossil forms; microstrobili borne among leaves on dwarf shoots; ovules on stalks borne among leaves; 1 extant genus.

DIVISION GNETOPHYTA
An artificial group containing 3 orders: Ephedrales, shrubs with reduced leaves and jointed stems; Welwitschiales, plants with a massive, fleshy stem bearing 2 large, leathery, strap-shaped leaves; and Gnetales, vines, shrubs, or trees with flattened angiospermoid leaves.

Critical appraisal. The classification presented above emphasizes that all gymnospermous plants are not closely related to each other. The characteristic of naked seeds was apparently derived among seed plants more than once. The fossil record indicates that the seed-fern–cycadophyte complex was separate from the conifer line from the very beginning. Plants called pteridosperms may not all belong together; Mesozoic forms were quite different from the Paleozoic ones. In fact, it cannot even be determined in some instances that seeds were actually borne on leaves in the Mesozoic forms. Glossopterids, in a sense, have seed-fern characters in that seeds were borne on fernlike leaves (although they were entire). The wood of glossopterids, however, is more like that of conifers.

Several fossil groups were not included in this classification because their relationships are still obscure. Some of these groups are in the orders Pentoxylales, Vojnovskyales, and Czekanowskiales.

(T.De.)

THE GYMNOSPERM DIVISIONS

Coniferophytes

The division Coniferophyta embraces living and fossil plants that usually have needle-shaped, evergreen leaves and seeds attached to the scales of a woody, bracted cone. Three English names—cedar, cypress, and pine—are each applied to unrelated kinds of conifers. Among living gymnosperm divisions, the coniferophytes show little similarity to the Cycadophyta and Gnetophyta but share several vegetative and reproductive traits with the Ginkgophyta. Coniferophytes are most abundant in cool temperate and boreal regions, where they are important timber trees and ornamentals, but they are most diverse in warmer areas, including tropical mountains.

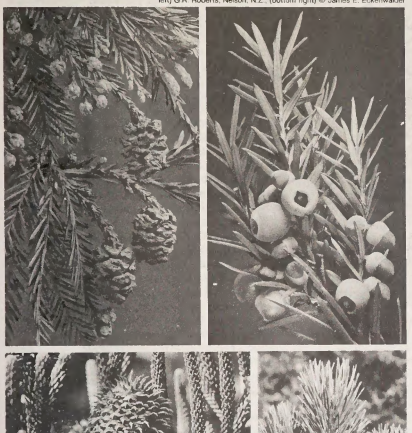
GENERAL FEATURES

Variety among conifers

Diversity of size and structure. The coniferophytes are the most varied gymnosperms (Figure 4). The world's oldest trees are the 4,900-year-old bristlecone pines (*Pinus longaeva*) of desert mountains in California and Nevada. The largest trees are the giant sequoias (*Sequoiadendron giganteum*) of the Sierra Nevada of California, reaching heights of more than 95 metres and weights of at least two million kilograms (4.4 million pounds; compared to 190,000 kilograms for the largest recorded blue whale). Wherever conifers grow, especially in temperate climates, one of these species is usually the tallest tree. In fact, the very tallest trees are the redwoods (*Sequoia sempervirens*) of coastal California, some of which are more than 110 metres tall.

The world's smallest trees probably are also conifers: the natural bonsai cypresses (*Cupressus goveniana*) and shore pines (*Pinus contorta*) of the pygmy forests (adjacent to the towering redwood forests) of the northern California coasts. On the sterile, hardpan soils of these astounding forests, the trees may reach full maturity at under 0.2 metre (7.9 inches) in height, while individuals of the same species on richer, deeper soils can grow to more than 30 metres. Other conifers, such as the pygmy pine (*Lepidothamnus laxifolius*) of New Zealand, the smallest conifer, are always shrubby and may mature as shorter plants (five centimetres in height) than the pygmy cypress, but with greater spread.

(Top left) John Markham—J. Allan Cash Photolibrary; (top right) Grant Heilman—Grant Heilman Photography, Inc., (centre left) © Gilbert Grant from the National Audubon Society Collection—Photo Researchers, (bottom left) G.R. Roberts, Nelson, N.Z., (bottom right) © James E. Eckenwalder



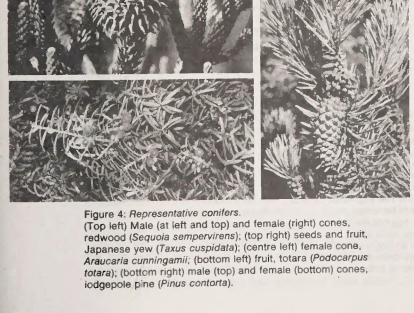


Figure 4: *Representative conifers.*
(Top left) Male (at left and top) and female (right) cones, redwood (*Sequoia sempervirens*); (top right) seeds and fruit, Japanese yew (*Taxus cuspidata*); (centre left) female cone, *Araucaria cunningamii*; (bottom left) fruit, totara (*Podocarpus totara*); (bottom right) male (top) and female (bottom) cones, iodgepole pine (*Pinus contorta*).

Distribution and abundance. Conifers almost cover the globe, from within the Arctic Circle to the limits of tree growth in the Southern Hemisphere. At these extremes, they often form pure stands of one or a few species. The immense boreal forests (or taiga) of northern Eurasia and North America are dominated by just a dozen species of conifers, with even fewer adjunct kinds of hardwoods. The richest north temperate conifer forests are those of mid-latitude mountain systems, where conifers also dominate in numbers. At lower latitudes and moderate elevations are found warm temperate woodlands and forests of pine (*Pinus*), oak (*Quercus*, a hardwood), and juniper (*Juniperus*), which vary in composition and density across Eurasia and North America.

Most tropical conifers are confined to cooler mountain areas where they form solid stands or grow with tropical hardwoods, while a few species inhabit lower elevations. The dammars (*Agathis*), for instance, dominate lowland tropical rain forests in Malaysia, Indonesia, and the Philippines, where they support an important forest industry. Conifers are widespread in southern temperate regions as well, generally with less dominance than in the north. Their greatest diversity is found in the humid portions of the three southern continents, but the largest areas are occupied by semiarid open woodlands of cypress pine (*Callitris*) in Australia and sederboom (*Widdringtonia*) in southern Africa.

Distribution

Conifer species are unevenly distributed. The Eurasian continent is richest in conifers, but every region has its own endemic genera and species. The most widely distributed genera are junipers (*Juniperus*) and pines (*Pinus*), both of which cover the northern continents and extend well into the tropics. Spruces (*Picea*) and firs (*Abies*) are only slightly more restricted. *Podocarpus* is the most widely distributed genus on the southern continents, followed by *Decussocarpus*.

At the other extreme, the most narrowly restricted endemic genera are *Austrotaxus*, *Neocallitropsis*, and *Parasitaxus* of New Caledonia, an island with the richest conifer flora in the world for its size (14 genera and 44 species). Other highly local genera include *Athrotaxis*, *Diselma*, and *Microcachrys* in Tasmania, *Fitzroya* and *Saxegothaea* in Chile, *Sequoiadendron* in California, and *Metasequoia* and *Pseudotaxus* in China. Most conifer genera fall between these extremes, with scattered distributions on one or more continents.

Uses in construction

Economic importance. Conifers provide all the world's softwood timber, the major construction wood of temperate regions, and about 45 percent of the world's annual lumber production. Softwoods have always had many general and specialty applications. The original great cedar (*Cedrus libani*) forests of the Middle East were felled to float the warring imperial navies of the ancient world. The same fate later befell the tall North American white pines (*Pinus strobus*) that masted the dominating British navies of the 18th and 19th centuries. Medieval archers drew longbows of the elastic yew wood (*Taxus baccata*). Victims of war and other dead in East Asia have been buried from earliest recorded times in coffins of sugi (*Cryptomeria japonica*) and sanmu (*Cunninghamia lanceolata*), relatives of the equally decay- and termite-resistant redwood (*Sequoia sempervirens*) and bald cypress (*Taxodium distichum*). In the family Cupressaceae are the fragrant cedars. Some are still used to line the chests that protect fine fabrics and furs against insects, but the wonderful fragrance of sharpened lead pencils has disappeared as eastern red cedar or pencil cedar (*Juniperus virginiana*) has been superseded by tropical hardwoods.

The domination of softwoods in lumber construction in northern temperate regions has been further extended by composite products such as plywood, particleboard, and chipboard. Other processed softwood products include paper and plastics derived from chemically treated wood pulp of spruces (*Picea*), tannins from the bark of hemlock (*Tsuga canadensis*), and naval stores (including turpentine) from many pines. Foods and beverages from conifers

include pine nuts and gin, which is flavoured with juniper berries. Canadian balsam, from resin blisters on the bark of the balsam fir (*Abies balsamea*) of northeastern North America, is used as a mounting medium for microscopic preparations.

Conifers are popular ornamentals in parks, cemeteries, and other public places, as well as around private homes and gardens. Although few species are grown indoors as houseplants, the traditional Christmas tree of western Europe and North America brings the fragrance and freshness of the forest into homes during the depth of the northern winter.

NATURAL HISTORY

All conifers share a typical seed-plant life cycle with a long-lived, dominant, photosynthetic, diploid sporophyte and a reduced, transient, dependent, haploid gametophyte (Figure 5). All phases of this general life cycle vary among conifers.

Sporophyte phase. The sporophytes of all conifers are trees or shrubs. They have a life span that ranges from a few decades to more than 4,000 years. The ecological role and way of life of this sole photosynthetic phase of the conifer life cycle varies with the size, form, and habitat of each species. Where conifers are ecologically dominant, as in the boreal and montane forests of the Northern Hemisphere, including the Douglas fir forests of western North America, they may make up 90 percent or more of all the living matter and they contribute greatly to the biosphere through photosynthesis.

Importance of fires

Fires play an important role in many conifer forests. Most conifers contain highly flammable resins. The flammability of these trees increases during hot, dry fire weather, when the water content of the living needles is drastically reduced. Few adult conifers can withstand a conflagration. The giant sequoia (*Sequoiadendron giganteum*) is an outstanding exception because it has insulating bark more than 50 centimetres thick.

Despite their susceptibility to fires, many conifers actually depend on these apparent catastrophes for regeneration.

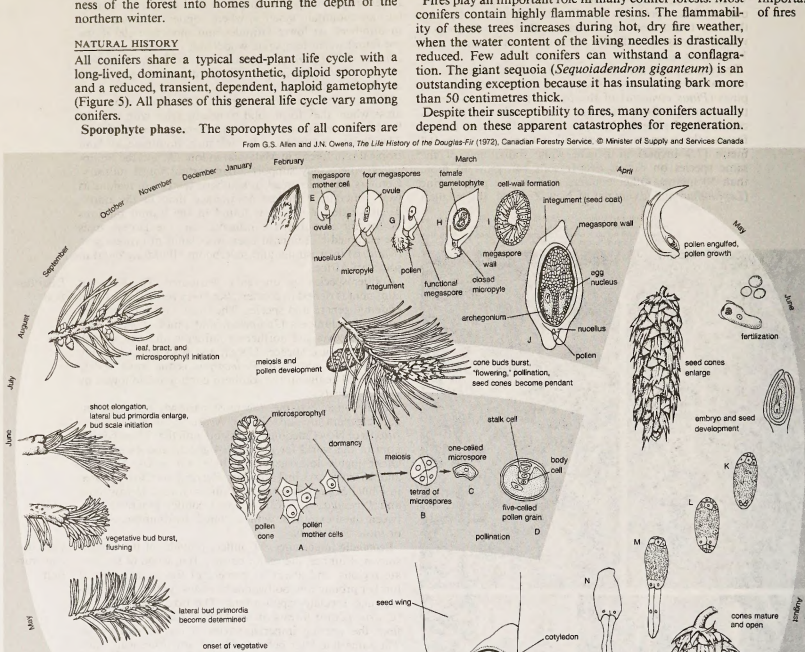


Figure 5: *Life cycle of the coniferophytes.*
(A) The pollen cone differentiates during summer and fall. Pollen mother cells begin meiosis in October; meiosis becomes arrested during winter and resumes in February. (B) Pollen mother cells form tetrads of single-celled microspores. (C) One-celled stage; microspore undergoes successive stages, culminating in (D) mature, five-celled pollen grains, which usually are shed early in April (pollination), after which pollen divides into two male cells (gametes). The stages of ovule and female gametophyte development occur during March, April, and May. (E) An ovule. (F) Meiosis occurs and four haploid megaspores are formed. (G) Three megaspores degenerate and one enlarges; pollination occurs. (H) The stigmatic tip grows inward, drawing pollen in; the megaspore divides without cell-wall formation. (I) Cell-wall formation after hundreds of nuclei have formed. (J) The large egg cell fills the archegonium before fertilization. (K) After fertilization, nuclei divide. (L) Cell walls form. (M) Primary suspensor cells elongate, forcing embryo (lower) tier through archegonial wall. (N) Elongation of suspensor cells and embryonal tube cells pushes embryo deep into female gametophyte tissue. (O) A section through the seed shows a dormant embryo embedded in the female gametophyte.

In such fire-dependent forests (including giant Sequoia groves, Douglas fir forests, boreal forests, low latitude pine forests, and Australian cypress pine woodlands), the dominant conifers are unable to regenerate among the more shade-tolerant species that grow up around them with time. Fires clear the understory to bare soil, stimulating the germination and establishment of seeds. Most of these species have cones that protect the seeds from the worst of the fire and then open to scatter them on the ash-fertilized seed bed. In the boreal forests of Eurasia and North America or in the pine forests of the southeastern United States, the cycle of fires and regeneration may repeat itself every 80 to 120 years, while the giant Sequoia is a fire-dependent successional species with individuals that can live for more than 3,000 years.

Contribution of flooding

At the other extreme are flooded swamp forests of bald cypress (*Taxodium*) in the southeastern United States and *shuaisuong* (*Glyptostrobus*) in southeastern China. Reproduction of these trees is as attuned to flooding as that of fire species is to scorched earth. Their seeds have air and resin pockets that allow them to float away to slightly raised areas revealed by receding floodwaters.

Without extremes of fire and flood, mesophytic species living in temperate and tropical mixed forests may dominate or grow scattered among other trees. The conifers with broad, flat blades rather than needle leaves almost all live in moist forests, as do most species whose seeds have fleshy structures that attract birds or small mammals.

Gametophyte phase. The gametophytes of conifers, like those of other seed plants, live out their brief, nonphotosynthetic lives almost entirely within the spore wall. All of their nutrition is derived from the parent sporophyte. The female gametophyte is never released from the tree until the seed matures. The male gametophyte is briefly separated from the sporophyte when pollen is released into the wind. These pollen grains contain an immature male gametophyte enclosed and dispersed in the microspore wall (Figure 5). In the Pinaceae, three successive divisions of the microspore produce a four-celled pollen grain within the microsporangium. It has two tiny prothallial cells (the last body remnants of the old free-living gametophyte), a tube cell, and a generative cell. After pollination, the tube cell develops the pollen tube and the generative cell divides to form a sterile cell and a spermatogenous cell. Prior to fertilization, the spermatogenous cell divides again to produce two male gametes. Other conifers share the later phases of male gametophyte development with the Pinaceae, but vary in the number of prothallial cells, from none in *Cephalotaxus, Sciadopitys,* Cupressaceae, and Taxaceae to as many as 40 in *Agathis* of the Araucariaceae, which has the most complex male gametophytes among the seed plants. Unlike the ovule (megasporangium), which houses a solitary female gametophyte, each microsporangium produces hundreds or thousands of pollen grains.

The pollen grain

The female gametophytes of conifers are more massive and complex than their male counterparts and basically resemble gametophytes of *Ginkgo* and the cycads. The life history of the female gametophyte begins with a protracted series of free nuclear divisions in the megaspore. At the end of these divisions, there may be up to 2,000 nuclei in a thin layer of cytoplasm pressed against the megaspore wall by a giant central vacuole. Cell walls then form between adjacent nuclei and gradually extend into the central vacuole until the entire gametophyte is filled with radially elongated alveolar cells that are equivalent to the prothallial cells of the pollen grain. This stage is followed by the appearance of archegonia at the micropylar end of the ovule. One to eight archegonia are usual in the female gametophyte of conifers, but there may be up to 200 in some species, each of which can produce an embryo if fertilized. Each archegonium has a single huge egg cell capped by a ventral canal cell and separated from the micropylar surface of the gametophyte by a short neck made up of one or two layers of neck cells. The archegonial end of the female gametophyte usually protrudes from the megaspore wall, which might otherwise prevent pollen tube penetration and fertilization.

Megaspore maturation

Pollination. Like *Ginkgo,* but unlike at least some cycads and gnetophytes, all conifers are pollinated by wind. Pollen may be produced in enormous quantities, particularly by species of true pine (*Pinus*), which can blanket the surface of nearby lakes and ponds with a yellow scum of pollen. The pollen grains of many Pinaceae and Podocarpaceae have air bladders, which orient them in a pollination droplet exuded by the ovules so that, when this droplet is withdrawn back into the ovule, the pollen tube will penetrate the nucellus to the archegonium. The pollen grains of families that lack prothallial cells are more or less spherical, lack air sacs, and can extend a pollen tube anywhere on their surface so that precise orientation is unnecessary. Some conifers lack a pollination droplet mechanism. Douglas fir pollen grains land on an enlarged, stigmalike growth of the micropyle, whence the pollen tubes grow into the nucellus and archegonium. The pollen grains of the Araucariaceae land on the scales of the female cone, and the pollen tubes reach the micropyle by burrowing into the cone scales.

Fertilization and embryogeny. After passing through the nucellus, the pollen tube presses between the neck cells of the archegonium and ruptures to release the tube nucleus, sterile cell, and the two male gametes (sperms). The ventral canal cell seems to help the male gametes enter the egg. One of the sperms fertilizes the egg nucleus to form the zygote, the first cell of the new sporophyte generation.

The zygote

The conifer zygote has fewer free nuclear divisions than do *Ginkgo* or the cycads. While many divide twice to form four free nuclei in the centre of the egg cytoplasm, there may be from zero to six free nuclear divisions. The nuclei usually move away from the micropyle, and cell-wall formation accompanies further cell divisions. The embryo develops and is fed by the nutritive tissue of the female gametophyte. The embryo rapidly enlarges at the expense of the maternal tissue and initiates typical sporophytic organization, consisting at maturity of a single axis with a root apex at one end and a shoot apex at the other, surrounded by two to eight cotyledons.

Germination. The mature seed consists of the dormant embryo embedded in remnants of the female gametophyte and megasporangium (nucellus) and surrounded by a seed coat. The seed coat of conifers is similar to that of other gymnosperms, developing from an integument with three distinct layers. Only the hard middle stony layer is evident in most conifers, protecting the embryo between seed release and germination. The outer fleshy layer is most prominent in those conifers, such as Cephalotaxaceae and some Taxaceae, whose seeds are dispersed by animals. The inner fleshy layer functions in the early development of the ovule, but persists only as a thin, papery membrane in the mature seed.

Germination proceeds immediately upon dispersal to a suitable site in many tropical conifers, but most cool temperate species require a winter period of cold, moist stratification before they will germinate. After the embryo absorbs water, a seedling root breaks through the seed coat and turns down into the soil.

The stem below the cotyledons elongates and lifts them above the ground. As the cotyledons begin to photosynthesize, they produce the energy needed for the early growth of shoots. The seedling shoot is densely clothed with needle-shaped juvenile leaves for a varying time until adult foliage forms; some cedars of the family Cupressaceae produce their first flattened side shoots within just a few nodes, while the longleaf pine (*Pinus palustris*) of the southeastern United States remains in a juvenile "grass" stage for years.

FORM AND FUNCTION

The basic organization of the conifer sporophyte resembles that of other seed plants. The four main organs (stems, leaves, roots, and sporangia) are all usually distinct from one another and have well-defined physiological functions. The considerable variation that occurs in these organs are commensurate with the varied environments in which the different species grow.

Stem function

Stem. Stems raise the photosynthetic leaves into the light and provide a channel for nutrients between the leaves and the roots. Most of the diameter of mature conifer

stems consists of secondary xylem (wood) produced by the vascular cambium, a permanent cylinder of dividing cells that lies just inside the bark. Because the growth of most conifers is cyclic, the wood generally consists of distinct growth rings. These are delimited by the juxtaposition of small dark cells from the end of a growing season with larger, lighter cells that mark resumption of growth. In temperate conifers, these rings correspond to annual growth flushes since no wood is formed during winter dormancy. Tropical species may lack this correspondence unless their habitat has strong seasonal variation in rainfall. Otherwise, they may have more than one growth ring in a year, each accompanying a new flush of leaves and branches.

The wood of conifers is generally more uniform and simpler in structure than that of flowering plants. One type of cell, the tracheid, serves both to transport water and to support the trunk so that conifers lack the more textured wood associated with the mixture of vessel elements and fibres in hardwoods. The wood may have longitudinal resin canals lined with living cells, but most of its living cells are found in the rays that extend horizontally from the centre of the stem to the vascular cambium. The pits, the tiny thinnings that connect adjacent wood cells, are quite varied among conifer families and genera and are one of the chief features used to identify conifer woods.

Bark

The bark that clothes the trunks may be thin, smooth, and flaky, peeling annually to reveal fresh bright colours, as in the lacebark pine (*Pinus bungeana*) of China and the tecate cypress (*Cupressus guadalupensis*) of southern and Baja California, or it may accumulate in broad, colourful plates, as in the ponderosa pine (*Pinus ponderosa*) of western North America, or as thick, fireproof, fibrous ridges on the giant sequoias (*Sequoiadendron giganteum*). The practiced eye can distinguish different species of co-occurring conifers by their bark alone.

Some conifers have transient special determinate twigs called short shoots that carry most of the photosynthetic leaves. In the bald cypress (*Taxodium*) and dawn redwood (*Metasequoia*), these short shoots look like double-sided combs and are shed each fall, to be followed by new ones in the same spots on the branches when growth resumes the following spring. The short shoots of pine (*Pinus*), the bundles of two to five (rarely one or eight) needles, are retained for up to 20 years but are finally shed and almost never grow out as branches. In contrast, the peglike short shoots of larch (*Larix*) and true cedar (*Cedrus*), like those of *Ginkgo*, are permanent, elongating slowly and producing new needles each year. The flattened sprays of branchlets of cedars in the family Cupressaceae may act like short shoots, falling intact after a relatively brief life, or they may provide the framework for further branching.

The strangest conifer shoots are the phylloclades that give the celery pine (*Phyllocladus*) its name. These flattened branches look like fern fronds, are green, and are the main photosynthetic organs of the tree. The true leaves are tiny scales that contribute little to overall food production.

Leaves. Leaves are specialized photosynthetic organs. The varied leaves of conifers are attached singly along the stems in a helical pattern or in opposite pairs or trios. Many Cupressaceae and a few other conifers have reptilian scale leaves only a few millimetres long. Diverse needle- and claw-shaped leaves range in length from about one centimetre in many conifers to more than 30 centimetres in some species of pine. Broad, flat, oblong blades up to 30 centimetres long occur in *Agathis* and some species of *Decussocarpus* from East Asia, and the monkey puzzle tree (*Araucaria araucana*) of Chile has hefty triangular wedges. The notched needles of the Japanese umbrella pine (*Sciadopitys verticillata*) are the oddest leaves among living conifers. They can be needlelike phylloclades or a pair of longitudinally fused needles. The largest coniferophyte leaves were those of the extinct genus *Cordaites*, with great paddle- or strap-shaped leaves up to one metre long and 15 centimetres wide.

Leaf function

Most conifer leaves, whatever their shape, minimize water loss. The reduced surface area of the scale- to needle-shaped leaves is an obvious example, but even the broader forms often have a thick, waxy coating that makes them waterproof. The gas-exchange openings of the leaves (stomates) are usually confined to a pair of narrow bands on the undersurface and are deeply sunken into chambers that separate them from direct contact with the dry air surrounding the leaf.

Roots. Roots gather water and mineral nutrients from the soil and anchor and support the above-ground portions. Most conifers have rather shallow, if wide-spreading, root systems, making the trunks highly susceptible to wind and surface disturbance. Even the largest conifers are no exceptions, and many of the individual giant sequoias (*Sequoiadendron giganteum*) in national parks in California are ringed by fences to reduce damage to the roots by the footsteps of millions of admiring visitors. The roots are the least-studied parts of the conifers but appear to be relatively uniform throughout the group. The specialized roots (called haustoria) by which the only parasitic conifer, *Parasitaxus usus*, attaches to the roots of its conifer hosts are an exception, but the oddest root structures are the "knees" of bald cypresses (*Taxodium distichum*), conical masses of woody tissue that emerge from the swamp waters around each tree. They probably help aerate the roots, which need oxygen to survive and function.

The fine feeding roots of conifers, like those of many flowering plants, do not work alone. They get a boost in their work by associating with specialized fungi whose structural filaments (hyphae) intermingle with them to form mycorrhizae. There are two distinct types of mycorrhizal associations among the conifers. The majority of species have vesicular-arbuscular mycorrhizae, called endomycorrhizae because the fungal hyphae actually penetrate the cells of the roots. All of the Pinaceae, and only the Pinaceae, have the other kind of root symbiosis, called ectomycorrhizal because the fungi sheath the rootlets and hyphae pass between the outer root cells without penetrating them. Each year, new roots grow out from the sheath and are recolonized only when the fungi later resume active growth. Ectomycorrhizal fungi reproduce through the attached mushrooms that are seen sprouting in pine forests, whereas endomycorrhizal fungi do so underground.

Associations

Strobili. The sporangia of vascular plants are technically asexual, but in the seed plants, because the gametophytes are wholly dependent upon the sporophyte and the female gametophyte even remains within the megasporangium, sexual terminology is extended to the sporophyte and sporangium-bearing organs. In all coniferophytes the organs containing microsporangia ("male") are separate from those bearing megasporangia ("female"), and in *Cephalotaxus*, some junipers (*Juniperus*), and the family Taxaceae these are found on different individuals.

The microsporangia of all conifers are attached to the scales of a simple pollen cone, or microstrobilus. The pollen cones usually consist of thin, parchmentlike scales (microsporophylls), each carrying two or more microsporangia on the lower surface. The number of scales and their size is quite variable, so that the overall length of the microstrobilus ranges from about two millimetres in some cypress (*Cupressus*) species to more than 20 centimetres in some *Araucaria* species.

Variations among megasporangia

Wide variations in the female (megasporangiate) reproductive structures among the conifers are the main basis for their classification. Most living conifers have a seed cone that is interpreted as a compound strobilus; each cone scale, inserted in the axil of a bract, is equivalent to an entire simple pollen cone. Fossil evidence shows how each ovule-bearing dwarf shoot of ancestral conifers was reduced and fused to form a single cone scale. Like the leaves, the bracts and scales are spirally arranged or occur in pairs or trios on the axis, and modern conifers have at least some fusion between each bract and its scale. The bracts and scales, or combined scales, vary in texture from woody to leathery, or even fleshy in bird-dispersed junipers (*Juniperus*) and the family Podocarpaceae. The number and size of cone scales varies widely among conifers, leading to seed cones that range from three millimetres long and less than one gram in *Microbiota* of the Amur region of Russia to more than 40 centimetres long in the sugar pine (*Pinus lambertiana*) of California

and more than 2.2 kilograms in some araucarians and the coulter pine (*Pinus coulteri*) of California.

The megasporangiate strobili of *Cephalotaxus* and most Podocarpaceae have the same basic structure as other conifer cones, but are so reduced and dominated by their much larger seeds that they do not look like cones. Even greater modification in the family Taxaceae has completely eliminated any trace of strobilar organization, and the solitary seeds sit at the tip of a short branch in a fleshy aril, a cup-shaped outgrowth of the seed stalk.

CLASSIFICATION

Distinguishing taxonomic features. Extant conifers differ from other gymnosperms in combining simple pollen cones with compound seed cones (or solitary terminal seeds in family Taxaceae). Although not possessed by all species, only conifers have needle leaves (of a variety of shapes) and pollen with bladders. Some other features, although not exclusive to coniferophytes, are also more common in them than in other gymnosperms. These include flattened, winged seeds (also in *Welwitschia*), scale-like foliage leaves (also in *Ephedra*), and the growth habit of a normal tree or shrub (also in *Ginkgo*).

Annotated classification. With more than 50 genera and 550 species, classification of the extant coniferophytes remains controversial. Disagreements exist throughout the classification, so that the numbers of orders, families, genera, and species are all disputed. The classification outlined here reflects current opinion for living conifers but simplifies extinct groups because the number of families to be recognized among the fossils is so uncertain. Extinct groups are indicated by a dagger (†).

†CLASS CORDAITOPSIDA

Paleozoic; strap-shaped leaves, up to 1 metre long, much larger than those of true conifers; both pollen and seed cones were compound and open, each bract with an axillary branch bearing numerous scale leaves surrounding pollen sacs or ovules; generally considered the ancestors of the Coniferales.

CLASS CONIFEROPSIDA

Contains coniferophytes; ovules attached to the scales of a condensed compound seed cone; families defined by seed-cone structure.

†Families Walchiaceae and Voltziaceae

Paleozoic and Mesozoic; show many stages in the transformation of the seed-bearing dwarf shoots of cordaiteans into the unified, flattened seed scales of modern conifers; foliage resembled that of araucarians; include *Walchia*, *Voltzia*, and *Voltziopsis*.

†Family Cheirolepidiaceae

Mesozoic; scales shed from the cone together with the seeds; large bracts remain attached to the axis in a semblance of a complete cone; distinctive pollen, called *Classopollis;* foliage resembles that found in the modern Cupressaceae; great variety of life-styles.

Family Pinaceae

Largest and most widespread and abundant modern conifer family in the Northern Hemisphere; woody, usually thin, cone scales carry two winged seeds and are fused to the bracts only at their bases; bracts usually hidden by the scales in mature seed cones but may be prominently exserted in some species; leaves are most often needlelike and spirally arranged, either singly or in clusters; pine (*Pinus*), spruces (*Picea*), firs (*Abies*), and larches (*Larix*) are all found across the Northern Hemisphere, while Douglas firs (*Pseudotsuga*) and hemlocks (*Tsuga*) are restricted to North America and Asia, *Cathaya, Keteleeria,* and *Pseudolarix* are restricted to China, and the true cedars (*Cedrus*) occur from Morocco to the Himalayas; 10 to 12 extant genera; about 200 species.

Family Araucariaceae

From Triassic; massive seed cones with a single large seed on each cone scale; highly reduced scales completely fused to the much larger bracts; species of *Araucaria* have branches densely clothed with prickly, spirally arranged, clawlike-to-wedge-shaped leaves; dammars (*Agathis*) have well-separated, oppositely arranged oval or oblong leaves, found in South America, Southeast Asia, and Australasia; 2 extant genera and 30 to 40 species.

Family Sciadopityaceae

Umbrella pine (*Sciadopitys verticillata*) usually included in the Cupressaceae, but recent work confirms its isolation from that family; seed cones superficially resemble those of the giant sequoia (*Sequoiadendron giganteum*), but the equal-sized scales and bracts fused for only about two thirds of their length, each having 5 to 9 seeds; foliage consists of whorls of about 15 to 20 double needles separated by stem segments with spirally arranged, nonphotosynthetic scale leaves; endemic to Japan.

Family Cupressaceae

Although species of this family are traditionally divided between two families, Cupressaceae for the cypresses (*Cupressus*) and similar genera and Taxodiaceae for the much more varied genera allied to the bald cypress (*Taxodium*) and redwood (*Sequoia*), present evidence shows that all belong to a single family; scales of seed cone intimately fused to the bracts; scale complexes vary in texture, shape, and arrangement on the cone; most have 3 to 5 seeds per scale, but the number ranges from 1 to about 20; leaves vary in shape from scales to clawlike or needlelike and are spirally arranged or in opposite pairs or whorls of 3; several genera, usually referred to as cedars (such as *Calocedrus, Chamaecyparis, Libocedrus,* and *Thuja*), have flattened sprays of frondlike branches closely covered with scale leaves; considerable diversity in both the Northern (18 genera) and Southern (11 genera) hemispheres; 50 or more species of junipers (*Juniperus*) are widespread, exceeding even the pines in their coverage of the Northern Hemisphere; other genera include *Athrotaxis, Callitris, Cryptomeria, Cunninghamia, Diselma, Fitzroya, Metasequoia, Microbiota, Neocallitropsis, Sequoiadendron,* and *Widdringtonia.*

Family Podocarpaceae

Seed cone is reduced, with 1 to few highly modified, brightly coloured, fleshy scales, each called an epimatium and surrounding a single seed; most with spirally arranged, yewlike needles, but scale leaves and opposite, broad, oblong blades (up to 30 cm long and 5 cm wide) are also found; about 130 species in 6 to 18 genera, including *Decussocarpus, Lepidothamnus, Microcachrys, Parasitaxus, Phyllocladus, Podocarpus,* and *Saxegothaea.*

Family Cephalotaxaceae

Seed cones highly modified with a few opposite pairs of small bracts, each with a greatly reduced scale remnant strongly dominated by a pair of ovules; only 1 ovule develops into a large seed with a fleshy seed coat; leaves are large yewlike needles carried in opposite pairs; found in East Asia, the plum yews (*Cephalotaxus*) are the second smallest and most local extant conifer family; 1 genus and 4 to 7 species.

Family Taxaceae

Solitary ovules borne at the end of a dwarf shoot bearing densely spiraled scale leaves; mature seeds surrounded by a fleshy aril and may also have fleshy seed coats; *Taxus* seed is highly toxic; pollen cones and flattened, needlelike leaves are more like those of other conifers than are their seeds; widely distributed in the Northern Hemisphere, while *Torreya* is found in restricted areas of both North America and East Asia, *Amentotaxus* and *Pseudotaxus* are localized in China, and *Austrotaxus* is endemic to New Caledonia; 5 genera and 16 species often segregated into an order separate from the Coniferales because of absence of seed cone. (J.E.Ec.)

Cycadophytes

Although some botanists prefer to restrict the term cycadophyte to the members of the division Cycadophyta, three groups of primitive seed plants are discussed here, of which the seed ferns (division Pteridospermophyta) and cycadeoids (division Cycadeoidophyta) are represented only by extinct forms. A third order, Cycadales (cycads), is today represented by 10 or 11 living genera and some 130–160 species.

The cycadophytes encompass a diverse collection of mostly extinct primitive seed plants that probably had their origins among the progymnosperms of the Devonian period (408 to 360 million years ago), possibly among a primitive, long-extinct group of non-seed-bearing plants, the Aneurophytaceae, in which disposition of fertile structures and patterns of branching bear some resemblance to those of seed ferns.

GENERAL FEATURES

Diversity. *Seed ferns.* A number of lines of seed-bearing gymnospermous plants are discernible among fossil plants of the late Paleozoic era (570 to 245 million years ago) and early to middle Mesozoic era (245 to 66.4 million years ago). Among them a rather loose assemblage of forms, referred to as seed ferns, or pteridosperms, is well represented. The Carboniferous period (360 to 286 million years ago) especially has been called the "age of ferns" because of the abundance of fossilized fernlike leaves.

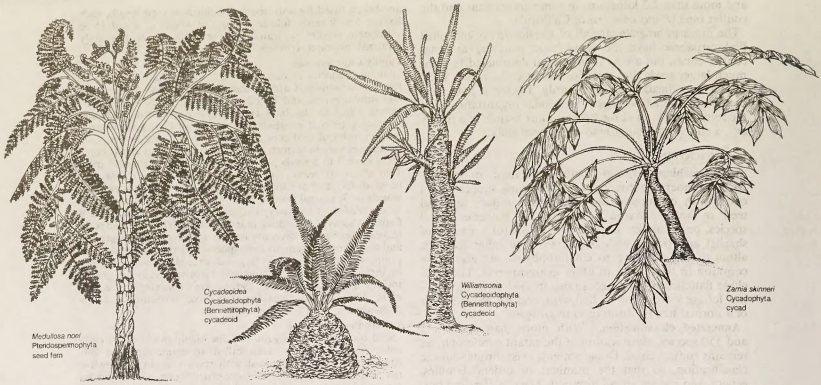


Figure 6: Representative cycadophytes.

From (*Medullosa noei, Cycadeoidea,* and *Williamsonia*) A.S. Foster and E.M. Gifford, Jr., *Comparative Morphology of Vascular Plants,* 2nd ed. (1974), copyright © 1959 and 1974 by W.H. Freeman and Co.; *Medullosa noei* redrawn from W.N. Stewart and T. Delevoryas, "The Medullosan Pteridosperms," *Botanical Review,* 22:45 (1956); and *Williamsonia* adapted from C.A. Arnold, *An Introduction to Paleobotany,* copyright © 1947 by McGraw-Hill, Inc., (*Zamia skinneri*) *Fairchild Tropical Garden Bulletin,* July 1987

In time, many of these "ferns" were recognized as seed plants, and it has been determined that seed ferns were a dominant vegetation in the late Paleozoic. Seed ferns generally are characterized as having been slender trees or, in some cases, woody, climbing vines, but generally with large, fernlike fronds.

Seed-fern foliage

Characteristic seed-fern foliage consisted of large compound leaves composed of second- and sometimes third-order branches (Figure 6). The latter bore fernlike leaflets, hence the name seed fern, although they are only remotely related to true ferns. Seed-fern stems generally possessed variable amounts of soft, loose wood and relatively large zones of cortex and pith; in this respect they resembled the stems of cycads and differed considerably from the stems of conifers, which have compact wood and relatively small zones of cortex and pith.

Reproductive organs of seed ferns were borne upon the foliage; single ovules and seeds were borne in place of pinnae, while male organs often occurred as compound pollen organs composed of partially or wholly united microsporangia. As in other gymnosperms, the ovule consisted of one megasporangium within a single integument. It is believed that, as the reproductive cycle progressed, the megasporangium, also called the nucellus, probably gave rise first to a quartet of megaspores. One of these then produced a large fleshy female gametophyte bearing several archegonia, each with a single egg. Following pollination and fertilization, the ovule developed into a seed with an embryo nested in the fleshy female gametophyte, which served as a food source during germination and seedling growth.

Cycadeoids. Although a few groups of pteridosperms persisted from the late Paleozoic era well into the Mesozoic, the common cycadophytes of the latter ages were members of the Cycadeoidophyta (also known as Bennettitophyta). They are well represented in the later Mesozoic era, well into the Cretaceous period (144 to 66.4 million years ago), by members of the genus *Cycadeoidea,* which had rather squat, barrel-shaped, unbranched trunks and once-pinnate compound leaves (Figure 6). The stems were armoured with the persistent bases of leaves; internally there was a thick pith surrounded by a narrow zone of vascular tissue from which vascular strands extended directly into the leaf bases. The fossilized trunks of these plants display scattered strobili among leaf bases of the characteristic armour. Fossil cycadeoids are widespread but are especially abundant in the Black Hills region of South Dakota.

Earlier in the Mesozoic era, cycadeoids of a more slender, branching form, exemplified by *Williamsonia,* were abundant (Figure 6). As in *Cycadeoidea,* the fronds were single pinnate compound leaves.

Cycadeoid features

The feature that set the cycadeoids apart from other cycadophytes was the compound strobili, which some, but not all, possessed. These strobili were composed of both male and female sporophylls, in some cases subtended by a system of bracts. Although often described as flowerlike and indeed sometimes depicted as having a floral, rosette form, cycadeoid "flowers," unlike true flowers (found in the angiosperms), were composed of sporophylls bearing "naked" (*i.e.,* gymnospermous) ovules. They are not now considered to have given rise to any group of the true angiospermous flowering plants.

Although cycadeoids flourished for millions of years, and must therefore be considered as a highly successful line of plants, they eventually became extinct in the Cretaceous period.

Cycads. The living cycads are for the most part palmlike, cone-bearing plants, generally of low stature (Figure 6). Although few genera, species, and individuals exist, they are extremely important plants in terms of the information that can be gained from studying them. Their reproduction is very primitive in that they rely on flagellated, motile male gametes (spermatozoids), a feature linking them with other plants fertilized by motile flagellated sperm (zooidogamous), such as ferns, club mosses, and other vascular cryptogams. Without knowledge of fertilization in the cycads and *Ginkgo,* it is highly unlikely that scientists would have more than remote theories as to the reproductive modes of seed ferns and other extinct groups of seed plants. Research on cycad reproduction is also providing information on the early origins of insect pollination, long thought to have evolved along with the relatively more recent angiosperms, or flowering plants.

Fossil remains

Distribution and abundance. Seed-fern fossils are found in both the Northern and Southern hemispheres, but many more have been described from Europe and North America than from other regions, primarily because many of the paleobotanical studies are concentrated there. Pteridosperms have been identified in Australia and India in recent years. In both hemispheres, seed ferns are common in coal measures, from which it may be inferred that, ecologically, they were plants of warm humid climates.

Abundant fossils of cycadeoids and cycads have been discovered and described from the Mesozoic era. The oldest remains of undisputed cycads date from the Triassic

period, 245 to 208 million years ago (*e.g., Leptocycas, Antarcticycas*), but some problematic forms (*e.g., Primocycas, Archaeocycas*) are of Paleozoic age. Most Mesozoic cycads resembled extant genera (*e.g., Cycadites, Pseudocycas, Cycadospadix*), and some are referred to present genera (*e.g., Macrozamia zamoides, Zamia coloradensis*). Fossil forms have been found in many places where they are now extinct (for example, Greenland, Antarctica, Alaska, Argentina, France, Austria), testifying to much milder climates in now temperate and even subarctic regions.

Ten genera of cycads are widely recognized. There are three endemic Australian genera—*Macrozamia* (14 species), *Lepidozamia* (two species), and *Bowenia* (two species); four American and Caribbean genera—*Microcycas* (one species), *Zamia* (about 35 species), and *Ceratozamia* and *Dioon* (10 species each); and two African genera—*Encephalartos* (about 40 species) and *Stangeria* (one species). The genus *Cycas*, with about 24 species, is the most wide-ranging, extending from eastern Australia westward across the Pacific and Indian oceans to Madagascar and the east coast of South Africa. In addition to the above well-known genera, a recent collection of cycad specimens from northwestern Colombia included a new genus now described under the name *Chigua*. *Chigua* reveals features hitherto undescribed in any American genus or species, for the specimens, which in most respects resemble *Zamia*, are unique in having leaflets with midribs and lateral veins, a characteristic formerly known only in *Stangeria*.

Ecology and habitats. Cycads are plants of subtropical habitats, where they occupy a variety of ecological situations ranging from rain forests, to mesophytic savannas, to near-desert scrublands. Now nowhere abundant in nature, wild populations of cycads in many regions are endangered. Only recently have Australian cycads been removed from the noxious weeds list (because of certain toxic properties dangerous to cattle) to a protected status.

NATURAL HISTORY

"Functional conifers"

Sporophyte phase. As in other gymnosperms, the large, woody plant is the sporophyte phase of the life cycle and typically is diploid in chromosome number. All cycads may be called "functional conifers," for all species bear strobili; these strobili are of a simple type, unlike those of true conifers, which bear more complex, compound strobili. It is not considered that this feature of cycads indicates anything other than a parallelism in evolution.

Cycad males and females are morphologically alike except for their sporophylls. Male sporophylls (microsporophylls) are spatulate organs bearing large pollen sacs (microsporangia) in clusters (sori) on their lower (abaxial) surfaces (Figure 7). Up to 200 cubic centimetres of pollen are produced by a single cone of *Cycas rumphii*, and some other species produce similar volumes. It was once estimated that one pollen cone of *Encephalartos* produced seven billion pollen grains having a total volume of about 300 cubic centimetres. While this enormous production would seem to be consistent with a system of wind dispersal, observations and controlled experiments strongly suggest that in most, or perhaps all, cycads, insect pollen vectors are necessary for effective pollination of ovules. The Mexican cycad *Zamia furfuracea*, for example, is pollinated by a small snout weevil, *Rhopalotria mollis*, which lays its eggs and completes its reproductive cycle in male cones. Emerging adults then carry pollen to female cones and pollination of ovules and subsequent fertilization of eggs occurs.

Gametophytes

Gametophyte phase. As in all other gymnosperms, male and female sporophytes of cycads produce, respectively, male and female gametophytes. The male gametophyte phase of the life cycle begins in the microsporangium with meiotic production of tetrads of microspores followed by the division of each haploid microspore into a three-celled pollen grain. Because it is multicellular, the pollen grain is considered to be an immature male gametophyte, but its further development into a sexually mature organism occurs only after it has been shed from the microsporangium and transported as a pollen grain to a megasporophyll—specifically, to an ovule, within which the male gametophyte grows to maturity.

Concurrently with pollen development, the ovule differentiates, and at the time of pollination it consists of a large megasporangium (nucellus) enclosed within a fleshy integument. At this time an opening at its distal end (the micropyle) permits pollen to enter the ovule. Over the next three to five months, the male gametophyte develops into a haustorial pollen tube, which eventually penetrates the nucellus and partially projects into an archegonial chamber.

Meanwhile in the nucellus, a single megaspore mother cell undergoes meiosis, forming a tetrad of haploid megaspores, only one of which survives to divide mitotically many times and form a large fleshy female gametophyte. The female gametophyte grows at the expense of nucellar tissue but remains enclosed within its remains. At its micropylar end, this gametophyte develops from one to many archegonia (commonly one to six in most cycads and up to 100 in *Microcycas*, only five or six of which are functional). Each archegonium is composed of a quartet of neck cells beneath which is a large egg. This egg is the largest known in the plant kingdom, being about three millimetres in length.

The development of male and female gametophytes is synchronized, and during the final week or so before fertilization, the male gametophyte forms large, multiflagellated spermatozoids. In all species but *Microcycas*, there is just one pair of sperm per pollen tube; in *Microcycas*, spermatozoids number about 12 to 16 per tube. What actually triggers sperm release into the archegonial chamber is unknown, but when it happens, the spermatozoids quickly move through the archegonial neck into the egg cytoplasm. One sperm loses its flagellature, and fusion of egg and sperm nuclei takes place. Subsequently, the zygote forms a single large embryo, other eggs meanwhile aborting. In the Florida cycad, *Zamia pumila*, the reproductive cycle occurs over a period of about 14 months, cones first becoming visible in October, pollination occurring in December, fertilization taking place in late May and early June, and embryogenesis and seed maturity being completed the following December. Similarly slow reproduction is typical also for other genera and species.

Seed development

As far as is known, cycad seeds have no dormant period or after-ripening requirements and in some cases actually begin germinating while still attached to sporophylls. Possibly some germination inhibitors are present in the outer fleshy layer of the seed, because its removal often accelerates germination, and treatment with scarifying agents also may enhance germinability. The germinating embryo remains attached to the female gametophyte for as long as two years, absorbing nutrition through its cotyledons, which remain embedded in the female gametophyte. The seedling rapidly develops a flesh taproot and root growth.

The outer layer of the seeds of *Cycas circinalis* and *C. rumphii* are thick and somewhat fibrous, and experiments which show them to be capable of long immersion in brine suggest that long-distance dispersal by ocean currents

(A) Adapted from A. Engler and K. von Prantl, *Die natürlichen Pflanzenfamilien*, Wilhelm Engelmann, Leipzig, 1918; (B) from A.S. Foster and E.M. Gifford, Jr., *Comparative Morphology of Vascular Plants*, 2nd ed. (1974), copyright © 1959 and 1974 by W.H. Freeman and Co.; *Ceratozamia mexicana* redrawn from A. Engler, *Syllabus der Pflanzenfamilien*, vol. 1, copyright © 1954 by Gebruder Borntraeger, West Berlin

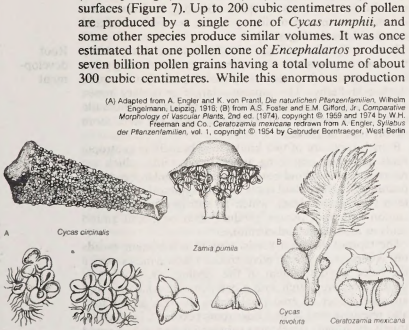


Figure 7: *Male and female reproductive structures of cycadophytes.*
(A) Microsporophylls and microsporangia of two cycads.
(B) Megasporophylls and ovules of two cycads.

may account for the presence of these species on remote Pacific islands. Little is known of natural seed dispersal of other cycads, but their bright red seed coats suggest a visual signal to animals. Mockingbirds, squirrels, and coatis are reported to disperse them. In general, however, cycad seeds, which are rather heavy, are poorly dispersed, and most germinate at the base of the parent, where they often languish and die.

FORM AND FUNCTION

Stem. Stems of cycads are characteristically short and stout, and while most genera have some species with subterranean, tuberlike stems, a majority of species are arborescent (Figure 6). The taller cycads include *Microcycas calocoma* (up to 10 metres high), *Macrozamia moorei* (up to 18 metres), *Dioon spinulosum* (up to 16 metres), *Lepidozamia hopei* (up to 18 metres), and *Encephalartos altensteinii* (up to 20 metres), but most of the arborescent (treelike) species have trunks only two to three metres high. The stems of most arborescent species are covered with an armour composed of the hardened leaf and cataphyll bases, but internally they are rather soft and fleshy, with a thick parenchymatous cortex, a large pith, and scanty woody tissue (Figure 8). In most cycads, the woody tissue is on the order of five to 10 millimetres wide, but *Dioon spinulosum* has an exceptional amount of wood, in some specimens up to 10 centimetres wide. This may constitute evidence of the primitive nature of the genus, because seed ferns also generally had stems with considerably more wood than those of most living cycads. Even in *Dioon*, there is no evidence of annual growth rings, so that age estimates must rely on other evidence, most often on counts of the whorls of leaf scars, which can be related to annual or biennial production of new leaf flushes. On this basis, it has been estimated that some cycads (notably *Dioon* and *Macrozamia*) may be as much as 1,000 years old; however, it is doubtful that most cycads are that old.

Stem size

Species of *Macrozamia, Encephalartos,* and *Cycas* often develop additional cylinders of vascular tissue, apparently formed from vascular cambia originating in the cortex. The result is a condition in which concentric rings of xylem and phloem are present, often two or three, but in exceptional cases, as many as 14. The xylem of cycad seedlings and that of some subterranean stems (*Stangeria, Zamia*) is composed of scalariform tracheids; in older stems, the tracheids exhibit primitive, multiseriate, bordered pits.

Another feature of those cycad stems in which terminal cones are produced is the presence of "cone domes" in the pith (Figure 8A). In longitudinal sections, the pith appears partitioned horizontally at intervals by vascular tissue. Each cone dome represents the displacement of a cone axis to one side as a result of the initiation and growth of the new vegetative apex.

Stem growth

The cycad stem grows from the tip (apically); the only lateral buds and branches are those unusually placed (adventitious) stems, whose buds arise by regeneration after the apical growth tissue (meristem) has been destroyed or as a result of wounding. Apical dominance and lack of branching bring about an apparent single-stemmed (monopodial) growth form, so that older plants become quite palmlike. This appearance, however, is deceptive, because in more than half the genera the apical meristem is converted from a vegetative to a reproductive function in that it is transformed into a strobilus (cone). A new vegetative meristem arises to one side of the cone meristem; subsequent growth and enlargement further displace the cone or cones to the side, so that the monopodial appearance is maintained even though the type of growth is actually sympodial (Figure 8). Only members of three genera (*Macrozamia, Lepidozamia, Encephalartos*) have cones initiated to the side and are truly monopodial; the remaining eight are considered sympodial.

Cycads have such thick stems that rearrangements of internal vascular connectives are not externally apparent. The cycad trunk is about as thick at its crown as at its base, thus furthering the resemblance to palms. Such stems, termed pachycaulous, result as in palms from activity of a primary thickening meristem (PTM) lateral to the apical meristem (Figure 8), which produces much greater increments of cortical parenchyma than would result if only an apical meristem were present. This is an important difference between cycadophytes and coniferophytes, for in the latter there is no PTM and the stem at its apical end is relatively smaller than at its base.

A further characteristic of cycad stems not occurring in cycadeoids, seed ferns, or coniferophytes is the presence of girdling leaf traces. In cacad stems, the vascular strands follow a circuitous route to the leaf bases, which is clearly seen in cross sections of stems. Girdling leaf traces are an important means of distinguishing between cycad and cycadeoid fossils (Figure 8).

Root development

Roots. Cycad seedlings initially form a stout, fleshy taproot that persists in subterranean forms for many years but is augmented by secondary roots which also are quite thick and fleshy. The taproots, larger secondary roots, and, in some cases, underground stems, have contractile elements in the pith and cortex that draw the stem more deeply into the ground.

Branch roots are of two kinds: long-branching geotropic roots and short-branching apogeotropic roots, which are referred to as coralloid because of their irregular, beady appearance. The coralloid roots contain symbiotic cyanobacteria (blue-green algae), which fix nitrogen and, in association with root tissues, produce such beneficial amino acids as asparagine and citrulline.

The taproot does not persist long in arborescent cycads but is replaced by large adventitious roots, which obscure the basic taproot system of the seedling. In all cycads, young roots are diarch with a parenchymatous cortex and an outer cover of epidermal scales. In this aspect they also resemble seed ferns. Older roots become triarch or tetrarch, eventually developing substantial amounts of wood and an outer covering of periderm.

Variation in leaf structure

Leaves. The leaves of cycads are for the most part once-pinnately compound; however, in the genus *Bowenia,* the

Adapted from (A,B,D) C.J. Chamberlain, *Gymnosperms: Structure and Evolution,* © 1935 by the University of Chicago Press; (B) after Dr. F. Grace Smith; (D) D.W. Stevenson, "Radial Growth in the Cycadales," *American Journal of Botany,* 67:468 (1980).

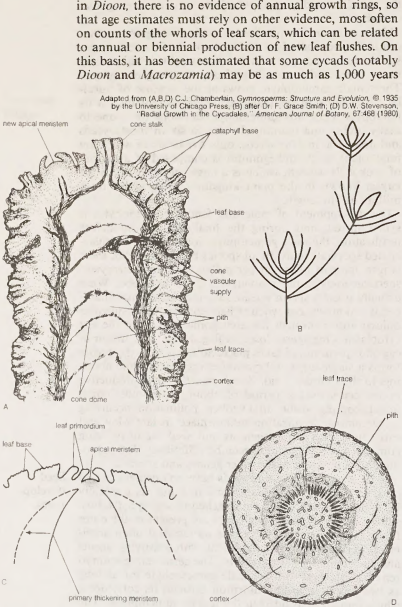


Figure 8: (A) Cone domes in a longisection of a *Dioon* stem. (B) Diagram of sympodial cycad stem architecture with axes much lengthened. (C) Diagram of a five-year-old stem apex of *Zamia,* illustrating lateral growth (arrow) resulting from action of a primary thickening meristem. (D) Cross section of a stem of *Zamia* showing girdling leaf traces.

leaves are bipinnate and quite fernlike. *Stangeria* also has fernlike leaves, and before cones were found to be associated with them the plant was described as a fern in the genus *Lomaria*. *Stangeria* leaves and those of the recently described *Chigua* are unique in possessing pinnately veined leaflets with midribs and side veins. *Cycas* pinnae also have midribs, but these lack side veins altogether. Pinnae of all other cycads have dichotomously branching, more or less parallel veins. The size of the cycad leaf is variable; *Zamia pygmaea*, the smallest cycad, has leaves about 20–30 centimetres long, while some species of *Macrozamia*, *Lepidozamia*, *Ceratozamia*, and *Cycas* have leaves three metres in length.

In cross section, the pinnae of most cycads are rather thick and sclerophyllous. The stomata are sunken and are of the type known as haplocheilic; that is, the guard cells arise directly from the mother cell, as contrasted with the syndetocheilic type, in which the guard cells are one division removed from the mother cell. The haplocheilic type is found in living conifers, pteridosperms, cycads, *Ginkgo*, and some others but not in the *Cycadeoidea*.

Sporophylls and strobili. Cycads are universally dioecious. Male plants produce pollen by leaf homologues called microsporophylls, and female plants produce ovules by leaf homologues known as megasporophylls (Figure 7). In all cycads, the microsporophylls are arranged spirally about a cone axis; in all cycads but *Cycas*, megasporophylls are similarly arranged. Megasporophylls of *Cycas* do not form a true cone but are arranged in two to three whorls at the stem apex. Later the stem resumes vegetative growth, and the megasporophylls then are interposed between whorls of foliar leaves and cataphylls; the usual arrangement is two to three whorls of leaves, then several whorls of cataphylls, followed by megasporophylls, but variations in this sequence are not unusual.

The megasporophyll of the Asian *Cycas revoluta* is considered to most typify the ancestral seed-fern condition. Each megasporophyll consists of a stalk, a fertile portion bearing two to six ovules, and an expanded terminal blade having fringelike "pinnae" (Figure 7). An evolutionary series of plant forms probably led toward the biovulate, peltate megasporophylls of such forms as *Encephalartos*, *Ceratozamia*, *Microcycas*, and *Zamia* (Figure 7). Microsporophylls similarly vary among cycads; those of *Cycas* are the more leaflike, those of *Zamia* less so (Figure 7). Microsporangia, which are found on the abaxial surface of microsporophylls, are usually numerous—several hundred in *Cycas*, several dozen in *Zamia*—and arranged in small clusters of two to five. They are the equivalent of sori of ferns and of pteridosperms. The cycad microsporangium resembles a clamshell, being somewhat flattened with an elongate suture (Figure 7).

CLASSIFICATION

Many botanists believe that extant gymnosperms represent at least two evolutionary lineages: one that leads to the extant conifers, taxads, and possibly *Ginkgo* and the Gnetales; and another that leads to the cycadophytes, represented today by seed ferns, cycadeoids, cycads, and perhaps others. Cycadophytes probably had their origins among the Devonian progymnosperms (Progymnospermopsida), although the particular antecedents are unknown.

DIVISION CYCADOPHYTA

Gymnospermous plants possessing compound leaves, ovules have 1 integument; seeds borne on either the foliage or megasporophylls; soft, loose wood contains scalariform tracheids and tracheids with multiseriate bordered pits; stem cross sections show wide zones of pith and cortex.

†DIVISION PTERIDOSPERMOPHYTA (seed ferns)

Primitive, primarily Paleozoic, primarily small trees or woody vines; large compound fronds; leaf-borne ovules; and microsporangia more or less united as synangia; subdivided into several groups: two orders, Caytoniales and Glossopteridales, persisted into Cretaceous, the latter sometimes included with pteridosperms, but commonly ranked separately and thought to be closely related to certain primitive angiosperms.

†DIVISION CYCADEOIDOPHYTA (BENNETTITOPHYTA)

Mesozoic; common and cycadlike; differ from cycads in having direct leaf traces, in sometimes being monoecious, and sometimes having bisexual cones.

DIVISION CYCADOPHYTA

Late Paleozoic? to the present; woody, coniferous plants with compound leaves, simple cones; flagellate motile male gametes; stout, fleshy stems; 4 families currently are recognized.

Family Cycadaceae

Generally restricted to species of *Cycas*; foliar, multiovulate megasporophylls arranged in an indeterminate strobilus; pinnae with a single midrib but lacking lateral, branch veins; 24 species defined.

Family Zamiaceae

Singly pinnate compound leaves, bearing leaflets with parallel, dichotomously branching veins (*Chigua*, if included, would be an exception); simple cones; female cones with biovulate megasporophylls; a total of about 112 species includes *Macrozamia*, *Lepidozamia*, *Ceratozamia*, *Encephalartos*, *Zamia*, *Microcycas*, and *Dioon*.

Family Stangeriaceae

Fernlike leaves bearing pinnae with a prominent midrib and numerous dichotomously branching lateral veins; simple cones; female cones with biovulate megasporophylls; includes only *Stangeria paradoxa*, a southern African cycad.

Family Boweniaceae

Differ from other cycads in possessing bicompound leaves; one genus, *Bowenia*, with 2 species. (K.J.No.)

Gnetophytes

The gnetophytes (division Gnetophyta) are characterized as a small group of vascular seed plants that are represented by only three genera: *Ephedra*, *Gnetum*, and *Welwitschia* (Figure 9). There are about 35 species in the genus *Ephedra*, 30 or more in *Gnetum*, but only one in *Welwitschia*. The three genera exhibit their great diversity in the immense variety of form and size among the various species.

GENERAL FEATURES

Most species of *Ephedra* are branched shrubs or small trees while others are vinelike, often clambering over other vegetation. Species are distributed in dry and cool regions in both the Eastern and Western hemispheres. In the Western Hemisphere, *Ephedra* occurs in desert areas in the southwestern United States, part of Mexico, and a wide area in South America. In the Eastern Hemisphere, certain species occur in China and in the Himalayas from Afghanistan to Nepal and Tibet at elevations of 2,700 to 4,350 metres (10,300 to 14,300 feet). *Ephedra*, known as

(Left) E.M. Gifford, (right) General Biological Supply House, Inc., Chicago, (bottom) Dr. C.H. Bonman



Figure 9: *Representative gnetophytes.*
(Left) Close-up of the shrub and the mature, fleshy seed cones of *Ephedra*. (Right) Foliage leaves and mature seeds of *Gnetum gnemon*. (Bottom) *Welwitschia mirabilis*, showing two gigantic leaves and terminal seed cones on a system of branches.

Importance

Ma-huang, has been a common medicine in China for thousands of years. The effective product, ephedrine, is prescribed for colds, to break a fever and induce sweating, and as a decongestant. Stem fragments of species in the southwestern United States and Mexico are used in the preparation of Mormon tea, Mexican tea, squaw tea, and desert tea. The drug ephedrine is now manufactured synthetically.

Most of the 30 species of *Gnetum* are lianas that climb high into trees of tropical rain forests in central Africa, Asia, northern South America, and islands between Australia and Asia. One species, *G. gnemon,* is a tree about nine metres tall. The leaves are large, much like those of many flowering plants. The seeds are eaten cooked or roasted, and young leaves also are eaten.

The most unusual and geographically restricted gnetophyte is *Welwitschia mirabilis,* which is unlike any other known plant in the world. It occurs in the Namib Desert of southwestern Africa near the coast of Angola and Namibia, as well as inland to about 150 kilometres. (Rainfall on the Namib Desert ranges from zero to 100 millimetres [four inches] per year). There are only two large, permanent, arching leaves; they split, and the tips die when they touch the hot sand. It is not clear how the plant obtains sufficient water to meet its needs. There is a taproot that may extend downward for one to 1.5 metres before it divides into numerous thin roots, which must tap a supply of water not available to other plants of the desert.

FORM AND FUNCTION

Stem. In *Ephedra* the stems are jointed, the basis for the common name joint fir. The leaves are arranged in pairs on the stem or in whorls of three with their bases forming a sheath around the stem at a node. The leaves of most species are reduced or scalelike, although in some species they may grow to three centimetres in length. The leaves of most species have two veins that are connected to two axial stem vascular bundles. The bulk of photosynthesis occurs in the green stems. Reduction in leaf surface may be related to the necessity of reducing water loss through the process of transpiration, a requirement for existence in a desert environment.

Leaves. The leaves of *Ephedra, Gnetum,* and *Welwitschia* are strikingly different in form and venation and provide morphological characters that are definitive for each of the genera.

The leaves of *Gnetum* resemble those of the angiosperms (the flowering plants) in form, structure, and venation. Two leaves at a node are broad and have a pinnate venation system (one midvein with lateral secondary veins that run to the leaf margin) and a meshwork of smaller veins. Older stems become hard by the production of wood (secondary xylem).

Welwitschia

A conspicuous vegetative feature of *Welwitschia* is the presence of two large permanent leaves. When a seed germinates, two seed leaves (cotyledons) emerge, followed by the production of the two permanent leaves. Soon after the development of the two leaves, growth activity shifts away from the tip of the stem (the apical meristem) to the bases of the two permanent leaves. Growth in this region takes place in a meristematic zone, which adds tissue at the base of each leaf. This development results in a bilobed crown and later to a circular concave disk surmounted by a band of meristematic tissue, which continues to contribute new tissue to the two large leaves. Thus, growth in *Welwitschia* has shifted from developing height to developing the two leaves outward. The leaves are perpetuated by the basal meristem at the rate of eight to 15 centimetres per year. The leaves become split and frayed in old plants. A leaf of one giant was reported to have an unbroken width of 1.8 metres and a length of 6.2 metres, of which 3.7 metres were of living tissue. Some plants are estimated to be 1,500 to 2,000 years old.

It is not known for certain how a plant with gigantic leaves such as *Welwitschia* can exist in a desert and carry on photosynthesis if the pores (stomata) of a leaf remain open during the heat of the day. It may be that *Welwitschia* has the type of photosynthesis in which stomata are open only at night. Carbon dioxide (CO_2) is stored in organic acids and then used during the day in photosynthesis. There are also about 100 days of brief morning fog. Water accumulates on the leaves and is thought to enter them, although the process has never been proved. More recently CO_2 has been shown to be taken up during the daylight hours through open stomata, leading to a tremendous loss of water by transpiration for the purpose of cooling the leaf surface.

One of the physical features that distinguish the gnetophytes from other gymnospermous divisions is the presence of vessels in the xylem (wood). A vessel is a longitudinal row of cells, called vessel members, which have several to many circular perforations in their end walls at maturity, providing an efficient pathway for the movement of water in the plant body. The possession of vessels is characteristic of the flowering plants (angiosperms) as well, and has led to speculation that gnetophytes, especially *Gnetum,* may have been close to the ancestral stock of some angiosperms.

Reproductive structures and function. The reproductive structures of gnetophytes are contained within strobili, or cones (Figure 10). Most species are dioecious; *i.e.,* one individual plant produces either pollen-producing or seed-producing cones, called microsporangiate and megasporangiate strobili, respectively. The two types of strobili are basically the same, consisting of oppositely arranged bracts in the axils of which are short fertile shoots.

Strobili and sporophylls

(Left) E.M. Gifford. (right) from A.S. Foster and E.M. Gifford, Jr., *Morphology and Evolution of Vascular Plants.* copyright © 1989 by W.H. Freeman and Co., redrawn from Sanwal, *Phytomorphology,* 12:243 (1962)

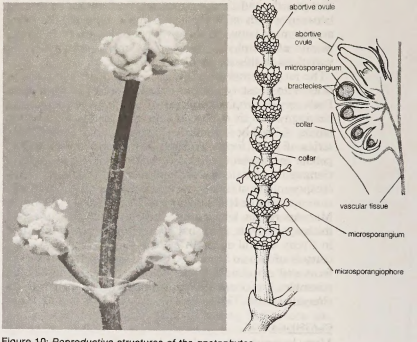


Figure 10: *Reproductive structures of the gnetophytes.* (Left) Pollen-producing strobili (cones) of *Ephedra.* The microsporangiophores extend out between the cone scales, and a pair of scalelike leaves appears at the nodes. (Right) Organization of the microsporangiate strobilus (cone) of *Gnetum gnemon* with (right) a longitudinal section through a node, showing four developing microsporangiate fertile shoots and one abortive ovule.

In *Ephedra* each fertile shoot of a pollen cone consists of an elongated modified structure, a microsporophyll or a microsporangiophore, which terminates in a cluster of sporangia, called microsporangia, that house the haploid reproductive cells (spores). The microsporangia are surrounded by a pair of bracteoles (scalelike leaves). Meiotic divisions in cells of the microsporangia produce the haploid pollen grains.

When shed from the microsporangia, the pollen grains (or male gametophytes) consist of five cells contained within the pollen-grain wall. The pollen grains are boat-shaped, with longitudinal ridges and furrows that run from one end of the grain to the other. Pollen of this type has been identified in the fossil record from the Permian (286 to 245 million years ago) to the present.

Megasporangiate, or seed-producing, strobili (female cones) consist of oppositely paired bracts in the axils of which are fertile shoots consisting of paired bracteoles

enclosing an ovule—the forerunner of a seed. The ovule consists of a delicate inner envelope, called an integument, that encloses a tissue (nucellus) in which a cell divides meiotically to produce a row of haploid cells called megaspores. One megaspore greatly enlarges and undergoes mitotic divisions, producing multiple nuclei that are not surrounded by walls. After 500 to 1,000 nuclei are produced, cell walls begin to form, converting the megagametophyte (or female gametophyte) into a cellular structure. At the upper end, egg-bearing protective structures called archegonia are formed, each of which contains a haploid egg cell.

Pollination and embryogeny. Just before the transfer of pollen grains to the ovule (pollination) some cells of the nucellus degenerate and a viscous solution is extruded through the opening (micropyle) of the now-extended integument. A pollination drop is formed at the tip of the integument. Transfer of pollen may be either by wind or insects. The pollination drop is very high in sugar content and attracts insects. As the sugary pollination drop dries, the pollen grains are pulled inward through the integument and come to rest on the nucellus. The pollen grain germinates, and a pollen tube enters the egg cell of the archegonium. A mature male gamete (sperm) fuses with the mature egg nucleus, bringing about fertilization. Eight or more diploid nuclei are produced and soon become surrounded by cellulose-containing walls. Each of these young embryos may develop, although only one generally reaches a fully developed stage in the mature seed. A seed consists of an embryo with two seed leaves (cotyledons), a stem axis, and a root, embedded in nutritive tissue of the female gametophyte. The pair of protective bracteoles become hard, and the seed is also surrounded by fleshy bracts that may become ivory, red, or orange in colour, perhaps an adaptation for animal dispersal. There appears to be no resting stage or dormant period for the seed, which may germinate immediately upon dispersal.

Fertilization

The strobili of *Gnetum* are compact or may become elongate with conspicuous nodes and internodes. In a microsporangiate, or pollen-producing, strobilus there are two fused bracts at each node forming a cup-shaped structure (collar) that encloses numerous fertile shoots. Each fertile unit consists of two bracteoles enclosing a microsporophyll or microsporangiophore with two sporangia at the tip. There may also be a top whorl of abortive ovules at each node.

The megasporangiate (ovulate) strobilus likewise consists of conspicuous nodes, each with a fused pair of bracts subtending several ovules. Each ovule comprises an integument and two ensheathing structures called envelops. In *Gnetum*, in contrast to *Ephedra*, fertilization occurs before the haploid female gametophyte becomes completely cellular. Nuclei become surrounded by membranes. A pollen tube pushes into the female gametophyte, and a male gamete fertilizes an egg. Several pollen tubes may enter the female gametophyte, resulting in more than one fertilization, although normally only one embryo develops to maturity. The seeds become fleshy and often are brightly coloured.

Reproduction in *Welwitschia*

In *Welwitschia*, branched reproductive shoot systems arise from the meristematic tissue at the bases of the two large permanent leaves. In both types of strobili, paired bracts and fertile shoots develop from the axils, as is true of the other two genera of gnetophytes. The fertile shoot of a microstrobilus has two pairs of bracteoles, the inner pair of which are fused, enclosing six microsporangiophores and a sterile ovule. In a megastrobilus the fertile shoot is made up of a pair of basal bracteoles and a pair of fused bracteoles surrounding an ovule.

At the time of pollination the integument of a fertile ovule in the female elongates and a pollination drop is formed at the tip. A cell in the nucellus of the female ovule undergoes meiosis, but no cell walls are formed between the four haploid nuclei. The four nuclei and their derivatives undergo repeated mitoses without cell-wall formation. Only later do cell walls form around several nuclei. No archegonia are formed. Subsequent development of the female gametophyte is without parallel in other gymnosperms or in angiosperms (Figure 11).

Some of the multinucleate female gametophyte cells grow into the nucellus of the ovule, into which the nuclei migrate. Pollen tubes, which grow downward in the nucellus from the male pollen grains after pollination, meet and fuse with the upwardly growing multinucleate tubes of the female gametophyte. "Fertilization bulbs" are formed, within which fusion takes place between a male gamete and one nucleus of the female gametophyte cells. A zygote is formed, and the young embryo then grows downward within the female gametophyte tube toward the cellular female gametophyte.

The later stages of embryo development take place within the tissue of the female gametophyte. A mature, ripe seed consists of a dicotyledonous embryo embedded in the female gametophyte and surrounded by papery fused bracteoles, constituting a "wing," which probably assists in seed dispersal.

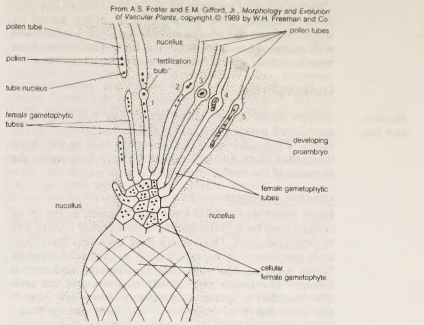


Figure 11: (Left to right) The stages in fertilization and proembryo development in *Welwitschia*. (1) Certain multinucleate female gametophyte cells form tubes that grow up into the nucellus. Pollen tubes growing down in the nucellus meet the upward-growing tubes of the female gametophyte. (2) Wall dissolution occurs, and fertilization takes place in the "fertilization bulb" to form a zygote (3). (4) The zygote divides, forming a two-celled proembryo. (5) The proembryo grows down inside the female gametophytic tube toward the nutritive tissue below.

CLASSIFICATION

Annotated classification.

DIVISION GNETOPHYTA

Longitudinal files of cells with pores in end walls (vessels) in xylem (wood); dioecious; strobili (cones) consist of bracts, in the axils of which are the reproductive structures, which have been likened to flowers of angiosperms.

Order Ephedrales

Shrubs to small trees; small leaves with 2 or 3 veins; mature cones often become fleshy and brightly coloured; 1 family, Ephedraceae; 1 genus, *Ephedra*, with about 35 species.

Order Gnetales

Mostly vines, but a few trees; large flat leaves that have reticulate venation; seeds may be brightly coloured; 1 family, Gnetaceae; 1 genus, *Gnetum*, with about 30 species.

Order Welwitschiales

Monotypic; 2 immense, permanent leaves, which become split and frayed with age; seeds with winglike extensions that may aid in dispersal; restricted to Namib Desert of Africa and vicinity; 1 family, Welwitschiaceae, with 1 member, *Welwitschia mirabilis*.

Critical appraisal. There have been attempts to demonstrate that the gnetophytes form a link between gymnosperms and angiosperms. Proponents of this concept cite the compound nature of both the pollen-producing and seed-producing strobili (some botanists have interpreted the strobili as "inflorescences"; *i.e.*, the axillary fertile shoots of both types of cones are considered to be flowers); and the presence of vessels in the xylem (wood),

a major difference between gnetophytes and all other gymnosperms. In recognition of these similarities, a classification has been proposed that placed the gnetophytes in the group Chlamydospermae, a group between the gymnosperms and the angiosperms. The word "chlamydospermae" means a seed with a "cloak," or envelope, a reference to pairs of bracteoles (scalelike structures) that surround a seed, similar to the ovary wall of the flowers of angiosperms. In another system the three genera are placed in one family in the order Gnetales.

More recently, the trend has been to establish a major group, the Gnetophyta (a division) or Gnetopsida (a class), consisting of three orders—Ephedrales, Gnetales, and Welwitschiales—each comprising one family and one genus. This system recognizes that there are certain shared characteristics but at the same time recognizes that there are significant differences between them.

The existence of gnetophytes in the fossil record has been based primarily upon the presence of pollen grains resembling those of *Ephedra* and *Welwitschia*. Leafy shoots with gnetalean features from the Early Cretaceous have been found associated with *Welwitschia*-type pollen.

Ginkgophytes

Maidenhair tree

Ginkgo biloba is also called maidenhair tree because of the bilobed nature of its leaves and its resemblance to the leaves of the maidenhair fern. *G. biloba* may be the oldest living seed plant, and it is regarded by some as one of the wonders of the world.

GENERAL FEATURES

Fossil leaves with similar form and venation to the living *Ginkgo* have been found in the Jurassic period (208 to 144 million years ago). These fossils have been described from such geographically separated areas as Australia, western North America, Mongolia, Alaska, England, and central Europe. The fossils vary greatly in form and are usually described as species of the genus *Ginkgoites*. Almost the same degree of variation in leaf form can be found on a living *Ginkgo* tree, however. Some paleobotanists, therefore, have recommended the abandonment of the genus *Ginkgoites* and the recognition of several species of *Ginkgo*.

There is one type of ginkgophyte leaf in the fossil record that is generally regarded as a distinct form and is given the generic designation *Baiera*. The leaf is deeply lobed into four segments and lacks a stalk (petiole). Following the Mesozoic era, *Ginkgo* declined progressively in its distribution, and some botanists believe that China was the last natural home of the maidenhair tree. Whether *Ginkgo* still exists in the wild state in China is unsettled. Although some evidence supports the view that the stands of trees in southeastern China are of a natural origin, some botanists contend that they may represent the offspring of trees cultivated in temple gardens for thousands of years. The latter may well be the case because, after the outer fleshy seed coat is removed, the seed kernel has been, and is still today, used as food in China and Japan. For 3,000 years or longer, extracts of the *Ginkgo* leaf have been recommended in Chinese medicine as benefiting the heart and lungs. *Ginkgo* has been investigated for its effects in the treatment of asthma, toxic shock syndrome, and various circulatory disorders.

Distribution of *Ginkgo*

FORM AND FUNCTION

Stem. At maturity a *Ginkgo* tree can reach heights of 20 to 30 metres (65 to 100 feet). Young trees often have a central trunk with regular, lateral branching; in older trees the branching is irregular.

A conspicuous feature of *Ginkgo* is the possession of long branches and short, or spur, branches. Leaves are produced on long branches during the spring growth. In subsequent years, clusters of leaves are formed on the lateral short branches. *Ginkgo* is deciduous, and the leaves of some varieties turn a beautiful golden colour in the autumn. The colour varies to some extent among horticultural varieties. There is some plasticity in growth form in that a short branch may become a long branch or the tip of a long branch may be converted to a short branch. The interplay between these two types of branches accounts for the more irregular shape of the older trees. Branching appears to be controlled by the distribution of auxin, a naturally occurring plant hormone.

The trunk diameters of the older specimens of *Ginkgo* may become large as a result of secondary growth. The vascular cambium gives rise to secondary phloem and secondary xylem (wood) for the conduction of water and dissolved minerals. The growth activity of the vascular cambium is sustained in the trunk and long shoots and produces a rather hard wood with well-defined growth rings. The activity of the vascular cambium persists in short shoots, but only a limited amount of soft wood is produced each year.

Leaves. One of the most distinctive features of *G. biloba* is the foliage leaf (Figure 12), which consists of a leaf stalk (petiole) and a fan-shaped dichotomously veined blade, or lamina. Although *biloba* (or bilobed) correctly describes the form of many *Ginkgo* leaves, there is a great range of variation in the degree of lobing and dissection among leaves of the same tree. Bilobed and undivided leaves occur on spur or short branches, while most of the leaves on the upper part of a long branch are divided by a deep sinus into two lobes, each of which is further dissected into segments. Multilobed leaves also occur on new branches (sucker shoots) arising from the tree trunk at ground level.

Variation in leaf lobing

(Top) Sonia Cook. (bottom) John H. Gerard

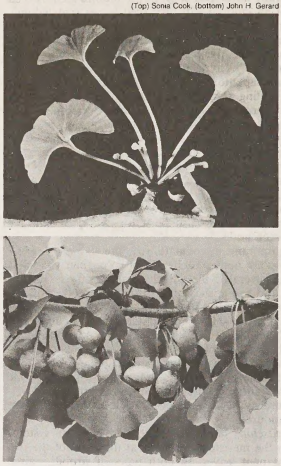


Figure 12: *Representative features of Ginkgo.*
(Top) Ovuliferous structures on spur branch. Note bilobed, fan-shaped structure of the leaf. (Bottom) Mature seeds attached to a short branch.

The dichotomous venation pattern in a leaf blade is a striking morphological characteristic of *Ginkgo*. Two vascular bundles extend through the petiole and give rise to two systems of dichotomously branched veins. This type of venation was also present in the leaves of extinct members of the Ginkgoales. Such a system of venation is often referred to as an open type, devoid of vein fusions. It has been shown, however, that vein unions may occur with some regularity.

Reproductive structures and function. Completion of the entire reproductive cycle, from the advent of pollination to the production of seeds with well-developed embryos, takes about 14 months. Pollination and the de-

velopment of the sexual, or gametophytic, phase of the life cycle occur in the first year (April to September), but embryo development is not completed until the spring of the following year.

Ginkgo is dioecious, which means that pollen-producing structures and ovules are produced on separate trees. The reproductive structures are restricted to the spur branches, where they are evident in the spring in the axils of bud scales and foliage leaves.

The pollen-producing strobilus is a loose, pendulous, catkinlike structure consisting of a main axis to which are attached numerous appendages, each of which usually bears two microsporangia at its tip. Meiosis occurs in cells of the microsporangia, giving rise to numerous haploid microspores. Cell divisions take place within the microspores, resulting in the formation of five-celled pollen grains (male gametophytes).

Ovuliferous structures

Ovuliferous structures (Figure 12) also arise in the axils of bud scales and the foliage leaves of spur branches. Each consists of a stalk that bears two or sometimes three or more erect ovules. An ovule is composed of an integument (the future seed coat) surrounding a tissue called the nucellus. It is in the nucellus that meiosis occurs, resulting in the formation of four haploid megaspore cells. It is at about this time that pollen grains are released from the microsporangia of male trees. The pollen (male gametophyte) is carried by wind currents and adheres to a pollination droplet, which exudes from the micropyle at the tip of the integument. Retraction of the droplet brings the pollen grains into a pollen chamber in the nucellus, where they develop into multibranched pollen tubes (male gametophytes).

One of the megaspores in the ovule that results from meiosis enlarges and undergoes a succession of free nuclear divisions (without wall formation). After about 8,000 haploid nuclei are produced, cell walls begin to form. After the female gametophyte becomes cellular, archegonia (normally two) are initiated at the surface toward the micropylar end of the ovule. An archegonium consists of neck cells and a large egg cell.

The basal end of the filament-like male gametophyte becomes suspended in a cavity above the female gametophyte (called the fertilization chamber). The spermatogenous cell of a male gametophyte divides, resulting in the production of two multiflagellated sperm. The sperm and the contents of the pollen tube are released into the fertilization chamber. The sperm swim in the liquid for a brief period of time. Approximately 1,000 flagella are attached to a spiral band at the anterior end. A sperm enters an archegonium and fuses with the egg nucleus. *Ginkgo* and the cycads are the only seed-producing plants that have motile sperm.

Fertilization

The growth of the embryo (embryogenesis) may begin shortly after fertilization but continues after the developing seeds fall to the ground. The embryo grows into the nutritive tissue of the female gametophyte. A seed at maturity (Figure 12) consists of a dicotyledonous embryo, nutritive tissue of the female gametophyte, and the seed coat, which is made up of a hard inner layer and a fleshy, orange-coloured outer layer. Because of the presence of butyric acid, upon decay the fleshy layer emits an odour similar to rancid butter.

CLASSIFICATION

In earlier systems of classification, *Ginkgo* was placed in the class Coniferopsida, along with conifers (*e.g.*, pine, fir, spruce). In recent years *Ginkgo* and its fossil allies have been placed in a separate group, the division Ginkgophyta (sometimes classified as the class Ginkgopsida), in recognition of the many characteristics outlined above. The early ancestral stock of *Ginkgo* extends back in the fossil record to a time coordinate with the ancestors of the conifers, but the two groups appear to have evolved independently.

DIVISION GINKGOPHYTA

Large trees; leaves typically fan-shaped and bilobed, or with more lobes, especially in fossil forms; leaves borne mainly on spur (short) branches; male and female trees; seeds with a fleshy outer layer that upon decay, emits an odour of rancid butter; a single order, Ginkgoales; a single family, Ginkgoaceae; a single extant genus, *Ginkgo*; 2 extinct genera, *Ginkgoites* and *Baiera*.

(E.M.G.)

BIBLIOGRAPHY. General works providing comprehensive coverage of the gymnosperms include K.R. SPORNE, *The Morphology of Gymnosperms: The Structure and Evolution of Primitive Seed-Plants*, 2nd ed. (1974), a compact summary discussing both living and extinct groups; THOMAS N. TAYLOR, *Paleobotany: An Introduction to Fossil Plant Biology* (1981), an excellent survey of fossil plants, including the history of the various gymnosperm groups and especially strong on the evolution of seeds; CHARLES JOSEPH CHAMBERLAIN, *Gymnosperms: Structure and Evolution* (1935, reprinted 1966), a classic description of the life history and morphology of all extant groups; W. DALLIMORE and A. BRUCE JACKSON, *A Handbook of Coniferae and Ginkgoaceae*, 4th ed., rev. by S.G. HARRISON (1966), a well-illustrated discussion of many representative types, including cultivated forms; and ERNEST M. GIFFORD and ADRIANCE S. FOSTER, *Morphology and Evolution of Vascular Plants*, 3rd ed. (1989), focusing on the structure and reproduction of vascular plants, including the gymnosperms.

For conifers, see GERD KRÜSSMANN, *Manual of Cultivated Conifers*, 2nd rev. ed. (1985; originally published in German, 1983), an extensively illustrated account of cultivars and species of conifers used in gardens; D.M. VAN GELDEREN and J.R.P. VAN HOEY SMITH, *Conifers* (1986), a pictorial work complementing Krüssmann's book with colour photographs; N.T. MIROV, *The Genus Pinus* (1967), a thorough treatment of all aspects of the biology of the many species of this important temperate genus; GEORGE S. ALLEN and JOHN N. OWENS, *The Life History of Douglas Fir* (1972), a detailed description of the reproductive cycle of the conifer; RUDOLF FLORIN, "Evolution in Cordaites and Conifers," *Acta Horti Bergiani* 15(11):286–388 (1951), a summary of the definitive research on the evolution of seed cones in coniferophytes, and "The Distribution of Conifer and Taxad Genera in Time and Space," *Acta Horti Bergiani* 20(4):122–312 (1963), an important survey, with thorough maps of distributions of living and fossil coniferophytes; CHARLES B. BECK (ed.), *Origin and Evolution of Gymnosperms* (1988), a collection of studies of fossil conifers with important discussions of early species; J.A. HART, "A Cladistic Analysis of Conifers: Preliminary Results," *Journal of the Arnold Arboretum* 68:269–307 (July 1987), a new approach to discovering relationships among conifer genera; and J.E. ECKENWALDER, "Re-evaluation of Cupressaceae and Taxodiaceae: A Proposed Merger," *Madroño* 23(5):237–256 (1976), an original look at family relationships among extant conifers.

A survey of cycadophytes is found in HAROLD C. BOLD, CONSTANTINE J. ALEXOPOULOS, and THEODORE DELEVORYAS, *Morphology of Plants and Fungi*, 5th ed. (1987). Other works include CHARLES JOSEPH CHAMBERLAIN, *The Living Cycads* (1919, reprinted 1965), the best comprehensive work on the cycads; CYNTHIA GIDDY, *Cycads of South Africa*, 2nd rev. ed. (1984), an excellent introduction to cycad morphology, noted for its beautiful colour illustrations of *Encephalartos* species in natural habitats; DIVYA DARSHAN PANT, *Cycas and the Cycadales*, 2nd ed. (1973), a fine presentation of all aspects of the life of *Cycas* and a valuable general reference to their anatomy and morphology; and PÁL GREGUSS, *Xylotomy of the Living Cycads, with a Description of Their Leaves and Epidermis*, trans. from Hungarian (1968), which emphasizes the structure of cycad xylem and includes important information on the habits and leaf morphology of cycads.

General treatments of the gnetophytes and ginkgophytes are found in PETER H. RAVEN, RAY F. EVERT, and SUSAN E. EICHHORN, *Biology of Plants*, 4th ed. (1986). WILSON N. STEWART, *Paleobotany and the Evolution of Plants* (1983), offers a more detailed account of the fossil record of *Ginkgo*; and CHRIS H. BORNMAN, *Welwitschia: Paradox of a Parched Paradise* (1978), is a brief study of *Welwitschia mirabilis*.

(T.De./K.J.No./J.E.Ec./E.M.G.)

The House of Habsburg

The House of Habsburg (also spelled Hapsburg; also known as the House of Austria) was one of the greatest of the sovereign dynasties of Europe.

ORIGINS

The name Habsburg is derived from the castle of Habsburg, or Habichtsburg ("Hawk's Castle"), built in 1020 by Werner, bishop of Strasbourg, and his brother-in-law, Count Radbot, in the Aargau overlooking the Aar River, in what is now Switzerland. Radbot's grandfather, Guntram the Rich, the earliest traceable ancestor of the house, may perhaps be identified with a Count Guntram who rebelled against the German king Otto I in 950. Radbot's son Werner I (died 1096) bore the title count of Habsburg and was the grandfather of Albert III (died *c.* 1200), who was count of Zürich and landgrave of Upper Alsace. Rudolf II of Habsburg (died 1232) acquired Laufenburg and the "Waldstätte" (Schwyz, Uri, Unterwalden, and Lucerne), but on his death his sons Albert IV and Rudolf III partitioned the inheritance. Rudolf III's descendants, however, sold their portion, including Laufenburg, to Albert IV's descendants before dying out in 1408.

AUSTRIA AND THE RISE OF THE HABSBURGS IN GERMANY

Albert IV's son Rudolf IV of Habsburg was elected German king as Rudolf I in 1273. It was he who, in 1282, bestowed Austria and Styria on his two sons Albert (the future German king Albert I) and Rudolf (reckoned as Rudolf II of Austria). From this date the age-long identification of the Habsburgs with Austria begins (see AUSTRIA). The family's custom, however, was to vest the government of its hereditary domains not in individuals but in all male members of the family in common, and, though Rudolf II renounced his share in 1283, difficulties arose again when King Albert I died (1308). After a system of condominium had been tried, Rudolf IV of Austria in 1364 made a compact with his younger brothers that acknowledged the principle of equal rights but secured de facto supremacy for the head of the house. Even so, after his death the brothers Albert III and Leopold III of Austria agreed on a partition (Treaty of Neuberg, 1379): Albert took Austria, Leopold took Styria, Carinthia, and Tirol.

King Albert I's son Rudolf III of Austria had been king of Bohemia from 1306 to 1307, and his brother Frederick I had been German king as Frederick III (in rivalry or conjointly with Louis IV the Bavarian) from 1314 to 1330. Albert V of Austria was in 1438 elected king of Hungary, German king (as Albert II), and king of Bohemia; his only surviving son, Ladislas Posthumus, was also king of Hungary from 1446 (assuming power in 1452) and of Bohemia from 1453. With Ladislas the male descendants of Albert III of Austria died out in 1457. Meanwhile the Styrian line descended from Leopold III had been subdivided into Inner Austrian and Tirolean branches.

The emperor Frederick III

Frederick V, senior representative of the Inner Austrian line, was elected German king in 1440 and crowned Holy Roman emperor, as Frederick III, in 1452—the last such emperor to be crowned in Rome. A Habsburg having thus attained the Western world's most exalted secular dignity, a word may be said about the dynasty's major titles. The imperial title at this time was, for practical purposes, hardly more than a glorification of the title of German king; and the German kingship was, like the Bohemian and the Hungarian, elective. If Habsburg was to succeed Habsburg as emperor continuously from Frederick's death in 1493 to Charles VI's accession in 1711, the principal reason was that the hereditary lands of the Habsburgs formed an aggregate large enough and rich enough to enable the dynasty to impose its candidate on the other German electors (the Habsburgs themselves had an electoral vote only in so far as they were kings of Bohemia).

For the greater part of Frederick's reign it was scarcely foreseeable that his descendants would monopolize the imperial succession so long as they did. The Bohemian and Hungarian kingdoms were lost to the Habsburgs for nearly 70 years from the death of Ladislas Posthumus in 1457; the Swiss territories, lost in reality from 1315 onward (see SWITZERLAND), were finally renounced in 1474; and Frederick's control over the Austrian inheritance itself was long precarious, not only because of aggression from Hungary but also because of dissension between him and his Habsburg kinsmen. Yet Frederick, one of whose earliest acts in his capacity as emperor had been to ratify, in 1453, the Habsburgs' use of the unique title of "archduke of Austria" (first arrogated for them by Rudolf IV in 1358–59), may have had some prescient aspiration toward worldwide empire for the House of Austria: the motto *A.E.I.O.U.*, which he occasionally used, is generally interpreted as meaning *Austriae est imperare orbi universo* ("Austria is destined to rule the world"), or *Alles Erdreich ist Österreich untertan* ("The whole world is subject to Austria"). He lived long enough to see his son Maximilian make the most momentous marriage in European history (see below); and three years before his death he also saw the Austrian hereditary lands reunited when Sigismund of Tirol abdicated in Maximilian's favour (1490).

Before explaining what the Habsburgs owed dynastically to Maximilian, mention can be made of a physical peculiarity characteristic of the House of Habsburg from the emperor Frederick III onward: his jaw and his lower lip were prominent, a feature supposed to have been inherited by him from his mother, the Mazovian princess Cymbarka. Later intermarriage reproduced the "Habsburg lip" more and more markedly, especially among the last Habsburg kings of Spain.

THE WORLD POWER OF THE HABSBURGS

Even before Frederick III's time the House of Habsburg had won much of its standing in Germany and in central Europe through marriages to heiresses. Frederick's son Maximilian carried this matrimonial policy to heights of unequalled brilliance. First he himself in 1477 married the heiress of Burgundy, Charles the Bold's daughter Mary, with the result that the House of Habsburg, in the person of their son Philip, inherited the greater part of Charles the Bold's widespread dominions: not the duchy of Burgundy itself, which the French seized, but Artois, the Netherlands, Luxembourg, and the County of Burgundy or Franche Comté. Secondly, though he failed after Mary's death in 1482 to secure Brittany also by a similar coup (France frustrated his proxy marriage to the Breton heiress Anne), he procured Philip's marriage, in 1496, to Joan, prospective heiress of Castile and Aragon: thus securing for his family not only Spain, with Naples–Sicily and Sardinia, but also the immense dominions the Spaniards were about to conquer in America. Maximilian's matrimonial achievements were the occasion of the famous hexameter *Bella gerant alii, tu felix Austria nube* ("Let others wage wars: you, fortunate Austria, marry").

Maximilian: the Burgundian and Spanish marriages

Since Philip I of Castile died prematurely, his son was already ruler of the Burgundian heritage and of Spain when, in 1519, he succeeded Maximilian as ruler of the Habsburgs' Austrian territories. In the same year he was elected Holy Roman emperor as Charles V.

Charles V

The threat of force as well as an enormous expenditure in bribes was necessary to secure Charles's election. Besides the fact that many of the German princes were reluctant to saddle themselves with so mighty a sovereign, there was the opposition of France, which saw itself already half-encircled, from the northeast clockwise to the southwest, by Charles's possessions. Dating from Maximilian's Burgundian marriage, antagonism between the French kings

and the Habsburgs was to persist, to the progressive detriment of the latter, until the middle of the 18th century, and until the second half of the 17th the other European powers would mostly sympathize with France. The Habsburgs in the 16th century were too formidable not to provoke envy and anxiety.

Charles V's responsibilities at the time of his becoming emperor were moreover too great for one man to assume, as he himself could acknowledge: they had to be divided. By the Treaty of Brussels (1522) he assigned the Habsburg-Austrian hereditary lands to his brother, the future emperor Ferdinand I. In 1521 Ferdinand had married Anna, daughter of Louis II of Hungary and Bohemia; and Louis II's untimely death in 1526, after his defeat by the Turks in the Battle of Mohács, prompted Ferdinand to stand as candidate for his succession, to which, despite rivals, he was elected.

The Habsburgs reached the zenith of their power before the end of the 16th century: the duchy of Milan, annexed by Charles V in 1535, was assigned by him to his son, the future Philip II of Spain, in 1540; Philip II conquered Portugal in 1580; and the Spanish dominions in America were ever expanding. There were, however, three faults in the power structure—two of them historical accidents, the

Table 1: The Imperial Succession of the Habsburgs, 1493–1740

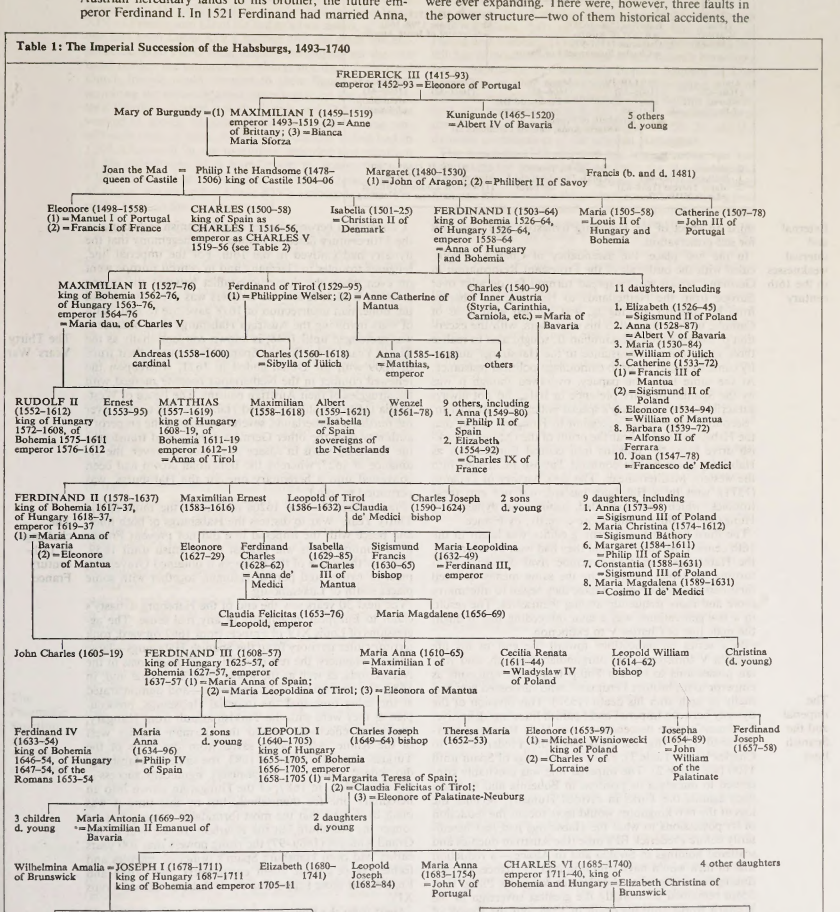


Table 2: The Spanish Habsburgs

CHARLES I of Spain (see Table 1) = Isabella of Portugal

- PHILIP II (1527–98), king of Spain 1556–98
  - (1) = Mary of Portugal
  - (2) = Mary of England;
  - (3) = Elizabeth of France;
  - (4) = Anna, dau. of Maximilian II
- Maria (1528–1603) = Maximilian II, emperor
- Joan (1535–73) = John, infante of Portugal

Children of PHILIP II:

- Charles (Don Carlos; 1545–68)
- 2 daughters:
  1. Isabella (1566–1633) = Albert of Austria
  2. Catherine (1567–97) = Charles Emmanuel I of Savoy

PHILIP III (1578–1621), king of Spain 1598–1621 = Margaret of Inner Austria

Children of PHILIP III:

- Anne (1601–66) = Louis XIII of France
- PHILIP IV (1605–65), king of Spain 1621–65
  - (1) = Elizabeth of France;
  - (2) = Maria Anna, dau. of Ferdinand III
- Maria Anna (1606–46) = Ferdinand III, emperor
- Charles (1607–32)
- Ferdinand (1609–41), archbishop of Toledo

Children of PHILIP IV:

- 2 children:
  1. Balthasar (1629–46)
  2. Marie Thérèse (1638–83) = Louis XIV of France
- Margarita Teresa (1651–73) = Leopold I, emperor
- CHARLES II (1661–1700), king of Spain 1665–1700
  - (1) = Marie Louise d'Orleans
  - (2) = Maria Anna of Neuburg

External and internal weaknesses in the 16th century

third an effect of the Habsburg dynasty's own measures for self-preservation.

In the first place, the ascendancy of Charles V coincided with the outbreak of the Protestant Reformation in Germany, which was to spread turmoil for decades over Europe from the Netherlands to Hungary. As Charles, from his Spanish upbringing, was imbued with ideas of Catholic uniformity and as his successors, with the exception of the enigmatic Maximilian II, sought also to realize those ideas, religious resistance to the Habsburgs' authority came to aggravate or to camouflage political resistance. At the same time, the papacy, overawed though it was by the Spanish military presence in Italy, did not always subscribe to the Habsburg's special policy for Catholicism.

Secondly, Ferdinand's accession to Hungary meant that the Habsburgs had to bear the brunt of the Ottoman Turkish drive from the Balkans into central Europe, just as Habsburg Spain had to confront Turkish incursions into the western Mediterranean. The great victory of Lepanto (1571), won by a Habsburg bastard, did not end these troubles, which were exploited, against the dynasty, by Hungarian dissidents and, more covertly, by France.

The third flaw in the Habsburg edifice was latent in the 16th century. Mindful of what they had won by marriages, the Habsburgs sought to preclude rival dynasties from turning the tables on them by the same means: to keep their heritage in their own hands, they began to intermarry more and more frequently among themselves. The result, in a few generations, was a fatal inbreeding that brought the male line of Charles V to extinction.

By a series of abdications toward the end of his life Charles V transferred his Burgundian, Spanish, and Italian possessions to his son Philip II and his functions as emperor to his brother Ferdinand, who succeeded him formally as such after his death (1558). This division of the dynasty between imperial and Spanish lines was definitive: Ferdinand's male descendants were Holy Roman emperors until 1740 (for the emperors from Frederick III to Charles VI see Table 1), Philip's were kings of Spain until 1700 (see Table 2). The imperial line was inevitably concerned to maintain its position in Bohemia and to assert itself against the Turks in divided Hungary, because the loss of the two kingdoms would have meant the reduction of its possessions to what the Habsburgs had had hereditarily before Frederick III's time (the Austrian duchies and scattered holdings in Swabia and in Alsace)—a reduction that in turn would have compromised its chances of continuing to be elected to the German kingship. Philip II of Spain remained territorially the greatest sovereign in the Western world until his death in 1598; but the Revolt of the Netherlands (see LOW COUNTRIES), which he proved unable to subdue, was an irritation that his English and French enemies did their worst to inflame.

The imperial and the Spanish lines

Cooperation between imperial and Spanish Habsburgs in the 17th century failed to maintain the hegemony that the dynasty had enjoyed in the 16th. For the imperial line, religious troubles in Germany and in central Europe went on even when the domestic conflict between the insane emperor Rudolf II and his brothers was over (1612); and the Bohemian insurrection of 1618 gave rise to that chain of wars involving the Austrian Habsburgs that, because it was prolonged until 1648, is known conventionally as the Thirty Years' War. For the Spanish Habsburgs, their truce of 1609 with the Dutch ended in 1621, whereupon the renewed conflict in the Netherlands became merged with the struggles of their Austrian cousins. The Peace of Westphalia (1648) finally abolished Habsburg sovereignty over the northern Netherlands, severely restricted the emperor's authority over the other German princes, and transferred the Habsburg lands in Alsace to France; however, the ordinance of 1627, whereby the Bohemian crown had been converted into a hereditary one for the Habsburgs, was permitted to stand.

The Thirty Years' War

France from the late 1620s had made the most of the Thirty Years' War to distress the Habsburgs of both lines, and peace with the imperial line did not prevent France from continuing its war against the Spanish until 1659, when by the Peace of the Pyrenees it obtained Gravelines, most of Artois, and part of Hainaut, together with some places south of Luxembourg.

The challenge of 17th-century France

The next 30 years saw the end of the Habsburg dynasty's claim to European hegemony in any real sense. The aggressions of Louis XIV of France, from 1667 onward, took territory after territory from the Spanish Habsburgs—large parts of Flanders, the rest of Artois, and other areas in the Netherlands, as well as the whole Franche Comté and, in 1684, the stronghold of Luxembourg—and demonstrated at the same time that the imperial Habsburgs, preoccupied as they were with the Turkish assault from Hungary, could not effectively defend the German frontier west of the Rhine. After being saved from the crisis of the Turkish siege of Vienna in 1683, the imperial Habsburgs did indeed obtain one dynastically significant success—the conversion, in 1687, of the Hungarian crown into an hereditary one for themselves—but by this time it was plain to Europe that the most formidable dynasty was no longer the Habsburg but the Bourbon. In the War of the Grand Alliance (1689–97) the rising powers that 100 years earlier had been Habsburg Spain's principal enemies and feeble France's most fluent encouragers, the Dutch and English, led those supporting the Habsburgs against Louis XIV.

Apart from the Bourbon ascendancy, there was a further reason for other powers to watch with jealous solicitude over the fate of Spain. The physical debility of Charles II of Spain was such that no male heir could be expected

The Spanish succession

to be born to him, and the question of his succession was one of great concern to the European powers. Up to 1699 it was understood that his crowns would pass to the electoral prince of Bavaria, Joseph Ferdinand, son of his niece Maria Antonia, daughter of the emperor Leopold I; and this arrangement was generally acceptable because, by transferring the Spanish inheritance to the Bavarian House of Wittelsbach, it would not necessarily upset the balance of power between the imperial Habsburgs and Bourbon France. In 1699, however, when Joseph Ferdinand died, the moribund Charles II's next natural heirs were the descendants (1) of his half-sister, who had married Louis XIV of France, and (2) of his father's two sisters, of whom one had been Louis XIV's mother and the other the emperor Leopold I's. Critical tension developed: on the one hand neither the imperial Habsburgs nor their British and Dutch friends could consent to their Bourbon enemy's acquiring the whole Spanish inheritance; on the other neither Bourbon France nor its British and Dutch enemies wanted to see an imperial Habsburg reunite in one pair of hands most of what the emperor Charles V had had in 1519. Charles II in the meantime regarded any partition of his inheritance as a humiliation to Spain: dying in 1700, he named as his sole heir a Bourbon prince, Philip of Anjou, the second of Louis XIV's grandsons. The War of the Spanish Succession ensued.

THE HABSBURG SUCCESSION IN THE 18TH CENTURY

To allay British and Dutch misgivings, Leopold I and his elder son, the future emperor Joseph I, in 1703 renounced their own claims to Spain in favour of Joseph's brother Charles, so that he might found a second line of Spanish Habsburgs distinct from the imperial; but when Joseph I died, leaving only daughters, in 1711, and was succeeded by his brother as emperor (Charles VI) and as ruler of the Austrian, Bohemian, and Hungarian lands, the British and the Dutch lost interest in making him king of Spain and together began serious negotiations with France. Their Treaties of Utrecht (1713), which recognized the Bourbon accession to Spain and to Spanish America, virtually forced the hand of the reluctant Charles, who made peace with France by the Treaty of Rastatt in 1714: out of the whole inheritance of the Spanish Habsburgs, he had finally to content himself with the southern Netherlands and with the former Spanish possessions on the mainland of Italy, together with Mantua (annexed by him in 1708) and Sardinia. Sardinia, however, was exchanged by him in 1717 for Sicily, which the peacemakers of Utrecht had assigned to the House of Savoy. With characteristic obstinacy, Charles remained technically at war with Bourbon Spain until 1720, when an armistice was declared (formal recognition of the Bourbon accession came only in 1725).

Meanwhile the extinction of the Spanish Habsburgs' male line and the death of his brother Joseph left Charles, in 1711, as the last male Habsburg. He had therefore to consider what should happen after his death. No woman could rule the Holy Roman Empire, and furthermore the Habsburg succession in some of the hereditary lands was assured only to the male line. In order, therefore, to secure the indivisibility of his Habsburg inheritance he issued his famous Pragmatic Sanction of April 19, 1713, prescribing that, in the event of his dying sonless, the whole inheritance should pass (1) to a daughter of his, according to the rule of primogeniture, and thence to her descendants; next (2) if he himself left no daughter, to his late brother's daughters, under the same conditions; and finally (3) if his nieces' line was extinct, to the heirs of his paternal aunts. The attempt to win general recognition for his Pragmatic Sanction was Charles VI's main concern from 1716 onward (his baby son died in that year). By 1738, at the end of the War of the Polish Succession (in which he lost both Naples and Sicily to a Spanish Bourbon but got Parma and Piacenza for the Habsburgs in compensation), he seemed to have won his point: Saxony, Bavaria (grudgingly and with an express reservation), Spain, Russia, Prussia, Hanover-England, and finally France (with a reservation about third-party rights) had all, in one way or another, acknowledged the Pragmatic Sanction. His hopes were illusory: less than two months after his death, in 1740, his daughter Maria Theresa had to face a Prussian invasion of Silesia, which unleashed the War of the Austrian Succession. Bavaria then promptly challenged the Habsburg position in Germany; and France's support of Bavaria encouraged Saxony to follow suit and Spain to try to oust the Habsburgs from Lombardy. Great Britain came, late enough, to support Maria Theresa rather out of hostility toward France than out of loyalty to the Pragmatic Sanction.

The Pragmatic Sanction and the Austrian succession

The challenge of Prussia

HABSBURG–LORRAINE

The War of the Austrian Succession cost Maria Theresa most of Silesia, part of Lombardy, and the duchies of Parma and Piacenza (Treaty of Aix-la-Chapelle, 1748) but left her in possession of the rest of her father's hereditary lands. Moreover, her husband, Francis Stephen of Lorraine, who in 1737 had become hereditary grand duke of Tuscany, was finally recognized as Holy Roman emperor, with the title of Francis I. He and his descendants, of the House of Habsburg–Lorraine (see Table 3), are the dynastic continuators of the original Habsburgs.

The peace of 1748 did not last long. Prussia was not satiated by the seizure of Silesia from the Habsburgs, and they in turn were even more determined to recover Silesia than anxious to ensure the protection of their outlying possessions in the Netherlands against the continuing danger of French attack. The so-called Diplomatic Revolution, which preceded the Seven Years' War of 1756–63, was the product, basically, of these situations: finding that their former British friends were more interested in conciliating Prussia than in abetting Austro-Russian plans for destroying it, the Habsburgs played their part in the "reversal of alliances" by achieving—without territorial profit—a reconciliation with France, hitherto their longest-standing enemy. An Austro-French entente was subsequently maintained until 1792: the marriage of the archduchess Marie-Antoinette to the future Louis XVI of France (1770) was intended to confirm it.

To secure their imperial status in Germany against Prussian enterprises, the Habsburgs exerted themselves to consolidate and to expand their central European bloc of territory. For this purpose Tuscany and the Netherlands were practically irrelevant. Tuscany in fact was kept separate from the ancient Habsburg inheritance: when the emperor Francis I died (1765), his eldest son, the emperor Joseph II, became coregent with his mother of the Austrian dominions, but Joseph's brother Leopold became grand duke of Tuscany; and similarly when Leopold succeeded to Joseph's titles (1790), his own second son succeeded to Tuscany as Ferdinand III. Thereafter the Tuscan branch of the Habsburgs remained distinct from the senior or imperial line.

The northeastward expansion of Habsburg central Europe, which came about in Joseph II's time, was a result not so much of Joseph's initiative as of external events: the First Partition of Poland (1772), which gave him Galicia and Lodomeria, was a Russo-Prussian arrangement disgusting to his conscientious mother, who remembered Silesia; and his subsequent acquisition of Bukovina (1775), geopolitically logical though it was as bridging a gap between his Transylvanian and his new Galician lands, was a side effect of the Russo-Turkish Treaty of Küçük Kaynarca (1774).

Joseph II was considerably more interested in westward expansion, over Bavaria, which would have both strengthened his western frontier strategically and enhanced his status among the German princes politically. Prussia's forceful opposition, however, reduced his gains in the War of the Bavarian Succession to the Innviertel (1779) and frustrated his plan for ceding the Netherlands to the House of Wittelsbach in exchange for Bavaria five years later (1784).

The French Revolutionary and Napoleonic Wars brought a kaleidoscopic series of changes. Three were clearly significant for the future of the House of Habsburg: (1) the formal dissolution of the Holy Roman Empire in 1806, in anticipation of which Leopold II's successor Francis II had in 1804 begun to style himself "hereditary emperor of Austria," a title that, as Francis I, he could retain come

The Austrian Empire

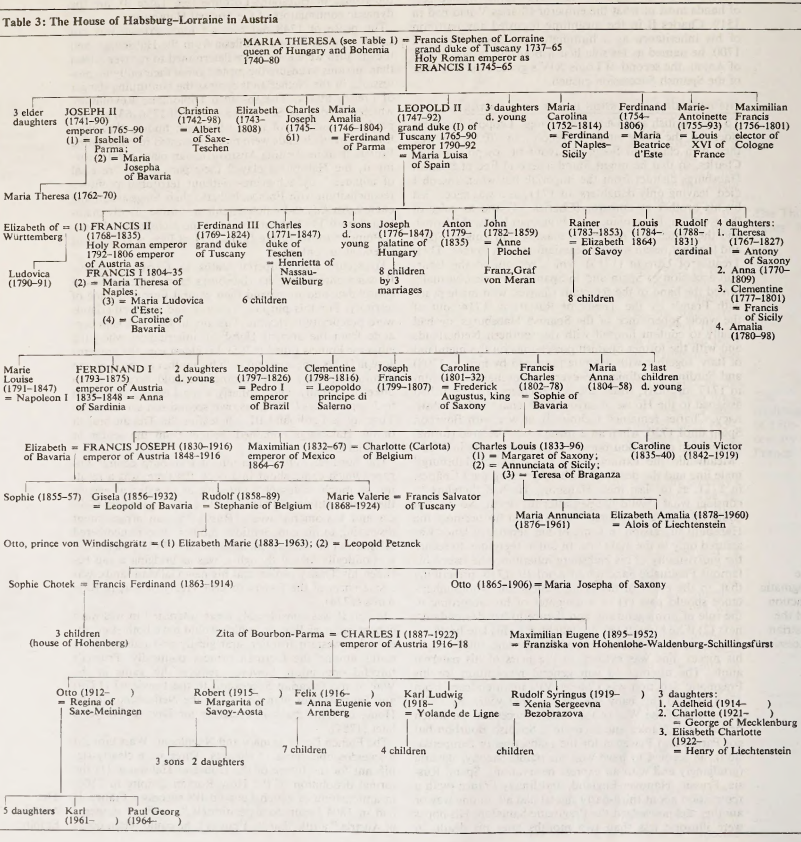
what might; (2) the definitive renunciation of the southern Netherlands by the Habsburgs in 1797; and (3) the awakening of the spirit of nationalism in the modern sense.

On Napoleon's downfall the Congress of Vienna (1814–15) inaugurated the Restoration, from which the battered House of Habsburg naturally benefitted. Francis I of Austria recovered Lombardy (lost in 1797), Venetia and Dalmatia (both of them acquired in 1797 but lost in 1809), and Tirol (lost also in 1809); Ferdinand III of Tuscany recovered his grand duchy; another Habsburg was recognized as sovereign duke of Modena, because his father, a brother of the Holy Roman emperors Joseph II and Leopold II, had in 1771 married the heiress of the House of Este; and Napoleon's Habsburg consort, Marie Louise, received the duchies of Parma and Piacenza for her lifetime (after which they were to revert to the Bourbons). The territory of Salzburg, which the Habsburgs had acquired in 1803 but lost to Bavaria in 1809, was finally restored to Austria in 1816. Though the Congress of Vienna did not restore Austrian rule over "Western Galicia" (the Habsburgs' share under the Third Partition of Poland in 1795, lost likewise in 1809), a small part of that area, namely the territory of Cracow, was annexed by Austria in 1846.

Nationalism against the dynasty

The history of the House of Habsburg for the century following the Congress of Vienna is inseparable from that of the Austrian Empire, a bastion of monarchical conservatism that the forces of nationalism—German, Italian, Hungarian, Slav, and Romanian—gradually eroded. The first territorial losses came in 1859, when Austria had to cede Lombardy to Sardinia–Piedmont, nucleus of the emergent kingdom of Italy, and could do nothing to prevent the same power from dispossessing the Habsburgs of Tuscany and of Modena. Next, the Seven Weeks' War of 1866, in which Prussia, exploiting German nationalism, was in alliance with Italy, forced Austria both to renounce its hopes of reviving its ancient hegemony in Germany and to cede Venetia. After this disaster the Habsburg emperor

Table 3: The House of Habsburg–Lorraine in Austria



Francis Joseph took a step intended to consolidate his "multinational empire": in 1867, to conciliate Hungary, he granted to that kingdom equal status with the Austrian Empire in what was henceforth to be the Dual Monarchy of Austria–Hungary. The result, however, was that the Magyars, jealous of their unique parity with the Germans and of their superiority over the non-Magyar peoples of their kingdom, rejected any suggestion of conciliating the Slavs and the Romanians of the Dual Monarchy by similar measures. The ardent German nationalists of the Austrian Empire, as opposed to the Germans who were simply loyal to the Habsburgs, took the same attitude as did the Magyars.

Remote from Austria's national concerns but still wounding to the House of Habsburg was the fate of Francis Joseph's brother Maximilian: set up by the French as emperor of Mexico in 1864, he was executed by a Mexican firing squad in 1867. No less grievous to the dynasty and of more concern to Austria–Hungary was the suicide of the crown prince Rudolf in 1889, though his fitness for the imperial and royal succession was questionable; and the scandalous misconduct of certain archdukes and archduchesses, in the imperial and in the Tuscan lines alike, further impaired the Habsburgs' personal prestige. The assassination of Francis Joseph's Wittelsbach consort Elizabeth in 1898 was to be followed in less than two decades by an assassination of far greater consequence.

In 1878 Austro-Hungarian forces had "occupied" Bosnia and Herzegovina, which belonged to decadent Turkey. In 1908 that territory had been formally annexed to Austria–Hungary, in a manner that was outrageous not only to Serbia (which coveted Bosnia for itself) but also to Serbia's patron, Russia. Visiting the Bosnian capital, Sarajevo, in 1914, the archduke Francis Ferdinand, heir presumptive to the Dual Monarchy (and incidentally legatee, from 1875, of the rights of the House of Austria–Este to Modena), was shot to death by a nationalist Serb. A month later the First World War was beginning.

World War I and after

World War I led to the dismemberment of the Habsburg Empire. While Czechs, Slovaks, Poles, Romanians, Serbs, Croats, Slovenes, and Italians were all claiming their share of the spoil, nothing remained to Charles, the last emperor and king, but "German" Austria and Hungary proper. On November 11, 1918, he issued a proclamation recognizing Austria's right to determine the future form of the state and renouncing for himself any share in affairs of state, and on November 13 he issued a similar proclamation to Hungary. Even so, he did not abdicate his hereditary titles either for himself or for the Habsburg dynasty. Consequently the national assembly of the Austrian Republic passed the "Habsburg Law" of April 3, 1919, banishing all Habsburgs from Austrian territory unless they renounced all dynastic pretensions and loyally accepted the status of private citizens. In Hungary, however, the collapse of the republican regime at the end of 1919 raised strong royalist hopes of a Habsburg restoration, and after the conclusion of the Treaty of Trianon (June 1920) Charles twice tried to return (March and October 1921). Under pressure from the other European powers, especially those of the Little Entente (Czechoslovakia, Yugoslavia, and Romania), the Hungarian parliament on November 3, 1921, decreed the abrogation of Charles's sovereign rights and of the Pragmatic Sanction.

Habsburg property rights in Austria, forfeited under the law of 1919, were restored in 1935 but withdrawn again by the German chancellor Adolf Hitler in 1938. After World War II the Allied Control Council in Austria in January 1946 declared that it would support the Austrian government in measures to prevent any return of the Habsburgs, and the law of 1919 was written into the Austrian State Treaty of 1955. In June 1961 the Austrian government rejected an application by the archduke Otto, head of the House of Habsburg, to be allowed to return to Austria as a private citizen, but in 1963 the administrative court of Austria ruled that Otto's application was legal. Because of Socialist opposition to his return, however, he was not granted a visa until June 1966 after the People's Party had won a majority in that year's general election. (J.R.-S.)

BIBLIOGRAPHY. For the English-language reader, the most comprehensive introduction to the subject remains WILLIAM COXE, *History of the House of Austria*, 3rd ed., 4 vol. (1847–53, reprinted 1901–05), an oft-neglected work covering Habsburg history from 1218 to 1848. ADAM WANDRUSZKA, *The House of Habsburg: Six Hundred Years of a European Dynasty* (1964, reprinted 1975; originally published in German, 1956), covers this same period in a brief but authoritative manner. Students able to read German will find a wealth of information in the multivolume series *Die Habsburgermonarchie, 1848–1918*, ed. by ADAM WANDRUSZKA and PETER URBANITSCH (1973–).

One of the best works on the early years of the Habsburg dynasty is ROBERT J.W. EVANS, *The Making of the Habsburg Monarchy, 1550–1700* (1979, reissued 1985). Works in English on the Habsburgs in the 19th and 20th centuries vary in scope and in quality. C.A. MACARTNEY, *The Habsburg Empire, 1790–1918* (1968, reprinted with corrections, 1971), also available in a condensed, more accessible version, *The House of Austria: The Later Phase, 1790–1918* (1978), provides a masterly survey by an expert. A.J.P. TAYLOR, *The Habsburg Monarchy, 1809–1918: A History of the Austrian Empire and Austria-Hungary*, new ed. (1948, reprinted 1976), is rather severely critical of the dynasty's failings.

Other studies include DOROTHY GIES MCGUIGAN, *The Habsburgs* (1966); VICTOR L. TAPIÉ, *The Rise and Fall of the Habsburg Monarchy* (1971; originally published in French, 1969); HUGH TREVOR-ROPER, *Princes and Artists: Patronage and Ideology at Four Habsburg Courts, 1517–1633* (1976, reissued 1991), an illustrated study; ROBERT A. KANN, *A History of the Habsburg Empire, 1526–1918* (1974); and JOHN LYNCH, *Spain Under the Habsburgs*, 2 vol., 2nd ed. (1981).

EDWARD CRANKSHAW, *The Fall of the House of Habsburg* (1963, reprinted 1983), together with his lavishly illustrated *The Habsburgs: Portrait of a Dynasty* (1971), makes a brilliant and easily readable vindication of the last emperor. ARTHUR J. MAY, *The Hapsburg Monarchy, 1867–1914* (1960), is fair and scholarly, with a good bibliography, and is continued in his *The Passing of the Hapsburg Monarchy, 1914–1918*, 2 vol. (1966). Z.A.B. ZEMAN, *The Breakup of the Habsburg Empire, 1914–1918* (1961, reprinted 1977), gives a just account of a subject usually misrepresented. Additional studies of the fall of the Habsburg dynasty include OSCAR JÁSZI, *The Dissolution of the Habsburg Monarchy* (1929, reissued 1961); and ALAN SKED, *The Decline and Fall of the Habsburg Empire, 1815–1918* (1989).

Genealogical detail is available in WALTHER MERZ, *Die Habsburg* (1896), which contains 19 tables; and MICHEL DUGAST ROUILLÉ, *Les Maisons souveraines de l'Autriche, Babenberg, Habsbourg, (Habsbourg-d'Espagne), Habsbourg-Lorraine, (Lorraine)* (1967), also well-tabulated and with illustrations. Discussions of physical heredity are found in OSWALD RUBBRECHT, *L'Origine du type familial de la maison de Habsbourg* (1910); WILHELM STROHMAYER, *Die Vererbung des Habsburger Familientypus* (1937); and JOHN LANGDON-DAVIES, *Carlos, the King Who Would Not Die* (also published as *Carlos, the Bewitched*, 1963).

There have been several good studies of economic development during the Habsburg reign. DAVID F. GOOD, *The Economic Rise of the Habsburg Empire, 1750–1914* (1984), is a solid survey. More advanced students may examine JOHN KOMLOS, *The Habsburg Monarchy as a Customs Union: Economic Development in Austria-Hungary in the Nineteenth Century* (1983). Also worthy of inspection is JOHN KOMLOS (ed.), *Economic Development in the Habsburg Monarchy in the Nineteenth Century: Essays* (1983). International relations and Habsburg foreign policy are the subjects of H.G. KOENIGSBERGER, *The Habsburgs and Europe, 1516–1660* (1971); F.R. BRIDGE, *The Habsburg Monarchy Among the Great Powers, 1815–1918* (1990); and SAMUEL R. WILLIAMSON, JR., *Austria-Hungary and the Origins of the First World War* (1991). (J.R.-S./Ed.)

Hamburg

The Free and Hanseatic City (Freie und Hansestadt) of Hamburg, on the Elbe River, is the second smallest of the 16 *Länder* (states) of Germany, with a territory of only 292 square miles (755 square kilometres). It is also the most populous city in Germany after Berlin and one of the largest and busiest ports in Europe. The official name, which covers both the *Land* and the town, reflects Hamburg's long tradition of particularism and self-government. Hamburg and Bremen (the smallest of the *Länder*) are, in fact, the only German city-states that still keep something of their medieval independence. The characteristic individuality of Hamburg has been proudly maintained by its people so that, in many spheres of public and private life, the city's culture has retained its uniqueness and has not succumbed to the general trend of standardization.

Hamburg, nonetheless, is a cosmopolitan city in its outlook. Although comparatively few foreigners live there, many pass through it. The city has dealings with a large number of nations, and it has more consulates than any other city in the world, except New York City. Shipping and trade have been Hamburg's lifeblood for centuries. Not surprisingly, its harbour has remained the city's most important feature.

Among Hamburg's many other facets are a network of canals reminiscent of Amsterdam; lakes, parks, and verdant suburbs full of gracious houses; elegant shopping arcades; richly endowed museums; and a vibrant cultural life. These are among the attractions that have contributed to a growing tourist industry. Although it was badly damaged during World War II, Hamburg has succeeded in maintaining a sense of old-world grace alongside its thriving commercial life.

This article is divided into the following sections:

Physical and human geography 472
The landscape 472
Site
The city layout
Architecture
Climate
The people 473
The economy 473
Industry
Trade
Transportation
Administration and social conditions 473
Government
Education
Cultural life 473
History 474
Early settlement and medieval growth 474
Evolution of the modern city 474
Bibliography 475

Physical and human geography

THE LANDSCAPE

Site. Hamburg stands at the northern extremity of the Lower Elbe Valley, which at that point is between five and eight miles (eight and 13 kilometres) wide. To the southeast of the old city, the Elbe divides itself into two branches, the Norderelbe and the Süderelbe; but these branches meet again opposite Altona, just west of the old city, to form the Unterelbe, which flows into the North Sea some 65 miles downstream from Hamburg. Two other rivers flow into the Elbe at Hamburg—the Alster from the north and the Bille from the east.

The city layout. The nucleus of the city is the Altstadt (Old Town), the former medieval settlement, bounded by the harbour and by a string of roads that follow the line of the old fortifications. Within this core there are few great buildings to remind the visitor of the city's thousand-year history apart from the five principal churches—Sankt Jacobi, Sankt Petri, Sankt Katharinen, Sankt Nikolai, and Sankt Michaelis—and none of these is in its original condition. Fire has destroyed almost all the older residences and warehouses, and what was left untouched by conflagration has often been rebuilt for contemporary purposes. There are, however, a few scattered survivals of older buildings. Moreover, the layout of the old city centre can still be detected in some of the ancient street names and in the *Fleete* (canals), which connect the Alster with the docks on the Elbe. One of the best views of the inner city is to be enjoyed from the Lombardsbrücke (Lombard Bridge), whence the towers of the five churches can be seen rising high against a skyline that is still relatively harmonious despite the presence of modern skyscrapers.

At the heart of Hamburg is a lake, measuring 455 acres (184 hectares), formed by the damming of the Alster and divided by the Lombardsbrücke into the Binnenalster (Inner Alster) and the Aussenalster (Outer Alster). Around the latter are elegant suburbs such as Rotherbaum, Harvesterhude, and Uhlenhorst. Many waterways, navigable by pleasure boats, run into the Aussenalster.

Architecture. The last intact ensemble of traditional Hamburg architecture is to be found in the Deichstrasse, one side of which backs onto the Nikolai canal. Its tall, narrow houses, resembling those of Amsterdam, were originally built from the 17th through the 19th century. It was in one of them, number 42, now a restaurant, that the devastating fire of 1842 broke out. Afterward the houses were rebuilt in the old style. Today the street is a protected area, and in recent years it has undergone extensive restoration. Many traditional restaurants are found there.

Another survival of older architecture is in the Krameramtswohnungen, near Sankt Michaelis. Consisting of two half-timbered brick buildings on either side of a narrow courtyard, it was built as a series of dwellings for the widows of shopkeepers and is the only surviving 17th-century construction of its kind in the city. Thoroughly restored between 1971 and 1974, it now forms a delightful secluded alleyway housing a restaurant, small shops, and a branch of the Museum für Hamburgische Geschichte (Museum of Hamburg History).

Of Hamburg's five great churches, the most imposing is probably Sankt Michaelis, an 18th-century Baroque-style Protestant church with a rich white-and-gold interior. It was destroyed by fire in 1906, rebuilt, devastated again during World War II, and restored yet again after the war.

The prosperous years 1890–1910 brought an abundance of fine architecture, examples of which can be seen in the spacious and elegant patrician houses around the Aussenalster. Many of these are now occupied by consulates. Another period of architectural flowering came in the 1920s and 1930s when there was a revival of the use of the traditional north German dark red brick as a building material, led by the architects Fritz Höger and Fritz Schumacher. A good example is Höger's Chilehaus, a massive office building constructed between 1922 and 1924.

More recently Hamburg has acquired its quota of starkly functional modern buildings, such as the Congress Centrum (Congress Centre; opened 1973) and the Fernsehturm (Television Tower), 271.5 metres (891 feet) high, but there is now a strong tendency to renovate old houses

rather than to demolish and build afresh. Thus the townscape of Hamburg as a whole has a human quality lacking in many German cities.

Climate. Hamburg has mild winters, late springs, relatively cool summers, high humidity, and frequent fog. The mean winter temperature is 34.2° F (1.2° C), and the mean summer temperature is 62.4° F (16.9° C).

THE PEOPLE

In 1965 the city had about 1,850,000 inhabitants, but since then the population has been slowly decreasing. More than three-fourths of the residents are Protestants, and the remainder are predominantly Roman Catholic. There is a small Muslim community, which includes many Turkish *Gastarbeiter* ("guest workers"), and the Jews, of whom there had been 27,000 in 1933 (when Hitler took power), now number only about 1,000.

THE ECONOMY

Industrial importance to Germany

Industry. Having absorbed Altona, Harburg, and Wandsbek in 1937, Hamburg has become Germany's major industrial city. All processing and manufacturing industries are represented there. Hamburg treats most of the country's copper supplies, and the Norddeutsche Affinerie, on Veddel, is Europe's second largest copperworks. The chemical, steel, and shipbuilding industries are also important, although shipbuilding has declined as a result of competition from Japan and Korea. Hamburg is also the most important centre in Germany, after Berlin, for newspaper and periodical publishing. Three nuclear plants, at Krümmel, Stade, and Brunsbüttel, provide power at a reasonable cost to the industries bordering the Unterelbe and to parts of Hamburg.

Trade. In the period of German partition, Hamburg handled more than half of West Germany's foreign trade, not only in the form of shipping cargo but also as rail and airfreight. Chief among imports are vegetable oils and fats, tea, coffee, petroleum, tropical fruit, and uncured tobacco. Exports include machinery, electrotechnical products, processed petroleum fuel and lubricants, copper, and pharmaceutical products.

The greatest economic centre of Germany, Hamburg since 1960 has become the site of first-class trade fairs. Many of the fairs and conventions are held at the Ernst-Merck-Halle exhibition grounds, located south of the Planten un Blomen park. An especially popular event is the international boat show, held each winter.

The port of Hamburg

Transportation. The harbour is Hamburg's "gateway to the world." More than 15,000 ships from over 100 countries pass through it each year. The city's Übersee-Zentrum is the world's largest roofed warehouse, and the Waltershof container terminal is the largest of its kind on the continent.

Harbour and city are well served by the German railway network, and the city has a good system of buses and underground trains. To relieve the central city from long-distance traffic, a tunnel was built (opened in 1977) under the Elbe as a part of the Stockholm–Lisbon highway.

The airport of Hamburg–Fuhlsbüttel, which dates from 1911, is one of the oldest in Europe. It has two runways from which even the largest jet-propelled aircraft can still take off.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. According to the constitution of June 6, 1952, legislative authority is vested in the Bürgerschaft (State Parliament), which comprises 120 members elected for a four-year term.

The Bürgerschaft elects the government, the Senat, which is organized on a collegiate basis; the president, or *erster Bürgermeister* ("first mayor"), is elected by the Senat itself annually, although in practice each incumbent normally stays in office at least four years. The Senat as a whole represents the Free and Hanseatic City of Hamburg in its dealings with the other federal *Länder*, with the federal government, and with foreign states. Each senator is responsible for a particular department, but administrative problems of a local nature are delegated to the district offices and to the local authorities.

The magnificent Rathaus (City Hall), where the Senat and the Bürgerschaft meet, in the centre of the city near the Binnenalster, was built late in the 19th century in the Neo-Renaissance style.

Hamburg's coat of arms displays a three-towered castle, intended to represent the Hammaburg, silver (argent), on a red (gules) field, its design being derived from the city's great seal of 1241. The state flag likewise shows a white castle on a red field. The state's anthem, "Stadt Hamburg an der Elbe Auen" ("City of Hamburg by the Meadows of the Elbe"), was written by G.N. Bärmann in 1828 and set to music by Albert Methfessel.

© Mike Yamashita—Aspect Picture Library



The Rathaus on the Alsterfleet in Hamburg.

Education. The Universität Hamburg, founded in 1919, is one of the largest in Germany, with some 46,000 students and faculties covering virtually every discipline except certain technological subjects. A second university, the Technische Universität Hamburg-Harburg, began classes in 1982. Hamburg also has state schools for music and interpretative art and for the sculptural arts, as well as some 250 research centres covering such areas as hydrography, oceanography, tropical medicine, shipbuilding, economics, meteorology, and particle acceleration.

CULTURAL LIFE

Among Hamburg's six principal museums, the Kunsthalle, founded in 1868 by Alfred Lichtwark, an outstanding patron of artists, is one of Europe's most remarkable galleries. It is particularly notable for its collection of 19th- and 20th-century works, including many of the German Romantic school. The Museum für Kunst und Gewerbe (Museum of Art and Crafts), founded in 1877 by the jurist Justus Brinckmann, has one of the most significant

collections of ancient artifacts in Germany and is also famous for its examples of Asian art and of Jugendstil (Art Nouveau). The Museum für Hamburgische Geschichte, which has grown from a collection of local antiquities started in 1839, contains a wide range of exhibits, from costumes to parts of old buildings and from architect's drawings to models of ships, shown in such a way as to present an impressive conspectus of the state's history. The Museum für Völkerkunde (Museum of Ethnology and Prehistory), founded in 1878, has impressive collections in its own fields. The Altonaer Museum, opened in 1863, specializes in north German subjects, with special attention to Schleswig-Holstein, and houses Germany's largest collection of old ships' figureheads. The Helms-Museum, in the Harburg district, is a local museum for the part of Hamburg south of the Elbe but also houses antiquities representing the prehistory and early history of the whole territory. The Ernst-Barlach-Haus, in Jenisch Park, was founded in 1961–62 by another great patron of the arts, Hermann F. Reemtsma, to make his private collection accessible to the public. Hamburg's once famous Zoological Museum was destroyed by bombs in 1943 after a century of existence.

Music and drama in Hamburg

The Hamburg Staatsoper, which dates from 1678, has won world renown. Its performances of classical and contemporary works bear comparison with those given by the great opera houses of Vienna, Milan, London, and New York City. The Deutsche Schauspielhaus, a leading theatre, enjoyed a particularly high reputation from 1955 to 1963, when Gustaf Gründgens directed and performed there. The Thalia-Theater, founded in 1843, with a multifaceted program that includes plenty of light entertainment, is popular with local audiences. All three establishments are generously funded by the state. The numerous other theatres include the tiny Piccolo-Theater and the Hansa-Theater, said to be the last genuine variety theatre in the German-speaking world. Plays of a local character or in Plattdeutsch (Low German) are performed in the Ohnsorg-Theater and sometimes in the Sankt Pauli-Theater, which dates from 1841 and is Hamburg's oldest playhouse.

The birthplace of Mendelssohn and Brahms, Hamburg has a sustained tradition of musical activity. Three great orchestras—the Philharmonische Staatsorchester, the Symphonie-Orchester des Norddeutschen Rundfunks, and the Hamburger Symfoniker—familiarize the public with classical and contemporary compositions. There are also groups specializing in chamber music, in choral performances, and in church music; and orchestras, choirs, singers, and instrumentalists from other parts of Germany or from abroad are also invited to Hamburg. The focal point of Hamburg's musical life is the Neo-Baroque Musikhalle, built in 1904–08 with money donated by the shipowner Carl Laeisz.

An important aspect of Hamburg's history is its prominence as a centre of newspaper and periodical publishing, which dates to the 17th century. By 1830 Hamburg had more newspapers than any other city in Germany. In the 1920s Berlin began its ascendancy as a press centre, and Hamburg fell into second place, which it still occupies. Its daily newspapers include the *Hamburger Abendblatt*, the *Hamburger Morgenpost*, and the *Bild-Zeitung*. In addition, a wide range of weekly and monthly magazines issues from the various publishing firms located in the city.

Sports

Sports have long been popular in Hamburg. The Hamburger Turnerschaft (1816) is Germany's oldest athletic club. The first German rowing club, founded in Hamburg in 1836, took part in 1837 in Germany's first official rowing race, against the English Rowing Club formed by members of Hamburg's then numerous English colony. The Hamburger Rennklub, for horse racing, was founded in 1852, and the North German Derby, first run in 1869, became an annual event, as the German Derby, from 1889 onward. Hamburg's first public football matches were played in 1881–82, after disputes about the rules of the game with the local Anglo-American Football Club. Another noteworthy event is the international tennis championship, which takes place every year in May.

An aspect of the city that cannot be ignored is its world-famous red-light district, centred on the Reeperbahn, where prostitution is legally permitted and kept under control by the police. Another famous Hamburg institution is the colourful Sunday-morning market held at the Fischmarkt by the harbour, where items ranging from fish to secondhand household goods are available.

History

EARLY SETTLEMENT AND MEDIEVAL GROWTH

Hamburg's history begins with the Hammaburg, a moated castle of modest size, built in about AD 825 on a sandy promontory between the Alster and Elbe rivers. In 834, during the reign of the emperor Louis the Pious, the castle's baptistery became the seat of an archbishopric, and Archbishop Ansgar made the young city of Hamburg the base of his missions to the heathens of northern Europe. Vikings burned the city in 845, and the rebuilt Hamburg was burned down again eight times in the following 300 years. By the end of the 11th century, Hamburg's role as the spiritual metropolis of the north was over, and henceforth commerce rather than religion was to be the principal *raison d'être* of the city. Between 1120 and 1140 some trading businesses were installed, and the foundation of Lübeck, on the Baltic, by Adolf II, count of Holstein, further promoted the economic development of Hamburg as Lübeck's port on the North Sea. In the autumn of 1188 a group of Hamburg entrepreneurs received from their feudal overlord, Adolph III of Schauenburg (Schaumburg), count of Holstein, a charter for the building of a new town, adjacent to the old one, with a harbour on the Alster River and with facilities for the use of the Elbe River as an outer roadstead. On May 7, 1189, the emperor Frederick I Barbarossa confirmed Count Adolph's dispositions in a charter granting special trading rights, toll exemptions, and navigation privileges.

In the 13th century Hamburg grew steadily in both area and economic importance, owing to the development of the Hanse (an association of merchants trading in a particular area) into a widespread association of north German merchant cities, the great Hanseatic League, in which Hamburg's role was second only to Lübeck's. A major entrepôt for the trade between Russia and Flanders, Hamburg proceeded to safeguard the trade routes by acquiring tracts of land along the branches of the Elbe in the immediate vicinity of the town and also on the estuary farther downstream (Ritzebüttel, nucleus of the later Cuxhaven, was acquired by Hamburg in 1393). It thus came to control the use of the river and to be recognized as the protector, in the emperor's name, of navigation on its lower course. Some political complications arose with the death, in 1459, of the last Schauenburg count of Holstein, since his princely rights in Germany passed thereafter to the royal house of Denmark, but Hamburg scarcely recognized Danish suzerainty in any but a formal way.

Role in the Hanseatic League

EVOLUTION OF THE MODERN CITY

Toward the end of the Middle Ages, the Hanseatic League gradually dissolved. Hamburg then went its own way and by 1550 had surpassed even Lübeck in economic importance. A stock exchange was founded in 1558 and the Bank of Hamburg in 1619; a convoy system for shipping was inaugurated in 1662, Hamburg's merchantmen being the first to be escorted on the high seas by men-of-war. About the same time marine insurance was first introduced into Germany. There were two causes for this new ascendancy: first, the wars of religion in the Low Countries in the second half of the 16th century had prompted many Dutch merchants to emigrate to the Unterelbe (Lower Elbe) region, with the result that Hamburg was henceforth to be the focus of their already established international commerce; second, the city had been so efficiently fortified in the decade 1616–25 that it could pursue its business untroubled throughout the worst crises of the Thirty Years' War (1618–48). By the end of the 17th century, Hamburg, with 70,000 inhabitants, was the largest city in Germany after Cologne.

The Treaty of Gottorp, concluded with the Danes on May 27, 1768, released Hamburg from theoretical subjec-

tion to the king of Denmark and so paved its way to being acknowledged, in 1770, as an "immediate" imperial city of Germany (that is, having no overlord other than the emperor). In addition, the treaty ceded to Hamburg the islands, from Veddel to Finkenwerder, that lay between the city and the left banks of the Elbe River and that, a century later, were to be the site of new docks. Hamburg, however, was not to enjoy its new advantage for long: the Napoleonic Wars overthrew the old order in Germany, and in 1810 the little state was annexed to Napoleon's French Empire.

After Napoleon's downfall (1814–15), Hamburg became a member state of the German Confederation, with the designation "Free and Hanseatic City of Hamburg" from 1819. Prosperity was quickly recovered, as Hamburg's trade was extended to newly opened territories in Africa, Asia, and the Americas. Even the great fire of May 1842, which devastated one-fourth of the city, did not check the booming economy, and the harbour was converted into one accessible at any time, without ships' having to depend on the state of the tides in the Elbe Estuary. Under the German Empire, founded in 1871, the political status of Hamburg was maintained, and development proceeded unchecked. The splendid Baroque houses of the densely populated Brook Island were demolished in the 1880s to make room for the warehouses of the new free port. By the end of the 19th century, in the course of which the population grew from 130,000 to 700,000, Hamburg had expanded far beyond its previous limits, absorbing such former suburbs as Sankt Pauli and Sankt Georg and spreading its tentacles into the countryside, toward Eimsbüttel, Eppendorf, Harvestehude, and Barmbek.

Hamburg entered the 20th century determined to maintain and to strengthen its position as "Germany's gateway to the world"; new docks and wharves were constructed on the left bank of the Elbe River. The outbreak of World War I in 1914 brought progress to a standstill, however. Hamburg's international trade collapsed, and its merchant fleet of 1,466 ships was virtually confined to port. After the war the victorious Allies demanded nearly all of Hamburg's ships by way of reparation from Germany.

The city between World Wars I and II

For many years after the war, Hamburg could undertake no further development because it had already exhausted all the potentialities of its territory. The Greater Hamburg Ordinance of January 26, 1937, changed this situation by allowing Hamburg to incorporate the neighbouring cities of Altona, Wandsbek, and Harburg, which until then had belonged to Prussia. The immediate prospect of expansion, with the development of these areas on a basis of large-scale planning, was shattered by the outbreak, in 1939, of World War II, during which repeated air raids demolished 55 percent of Hamburg's residential area and 60 percent of the harbour installations and killed 55,000 people. When the war ended in 1945, only the most strenuous efforts could supply the elementary needs for Hamburg's survival.

Reconstruction proceeded rapidly, however. Symptomatic of the city's postwar commercial efflorescence was the vast new business district City-Nord, built in the 1960s. At the same time, nightclubs on the Reeperbahn became proving grounds for British rock and roll bands—most notably the Beatles—who took advantage of a direct ship route from Liverpool, England. In 1962 the city experienced a flood, which killed more than 300 people and destroyed much of the old part of the city. In the mid-1960s, Hamburg's population exceeded 1,800,000, though it has fallen in the decades since, owing to a population shift toward the suburbs. With continued immigration of foreigners to the city, Hamburg's foreign-born population reached 10 percent by the 1980s.

The unification of Germany in 1990 increased trade between the city and eastern and central Europe. During the 1990s the city underwent continued modernization. In 1993 Hamburg hosted a multimedia festival, an exhibition on the use of modern communication methods in business and the arts. The following year the city became the seat of a Roman Catholic bishopric.

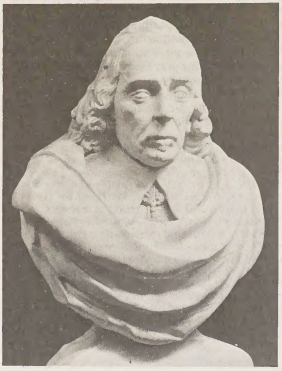
Hamburg's cherished traditions, together with its thriving business and cultural life and the energy of its inhabitants, make it one of the most vibrant cities in the world.

BIBLIOGRAPHY. BERNHARD MEYER-MARWITZ, *Das Hamburg Buch* (1981), a concise and lively book for the general reader; DAVID RODNICK, *A Portrait of Two German Cities: Lübeck and Hamburg* (1980), the city's history and social conditions; ECKART KLESSMAN, *Geschichte der Stadt Hamburg* (1981), a comprehensive and detailed study; HEINRICH REINCKE, *Hamburg: Ein Abriss der Stadtgeschichte von den Anfängen bis zur Gegenwart* (1926), a historical presentation; WILSON KING, *Chronicles of Three Cities—Hamburg, Bremen, Lübeck* (1914), an informative illustrated source; MARTIN CAIDIN, *The Night Hamburg Died* (1960), dealing with the events of World War II; and VOLKER PLAGEMANN (ed.), *Industriekultur in Hamburg* (1984), a well-illustrated symposium focussing on social and cultural aspects of the city's industrial history. The architecture of the city is explored in *Die Bau- und Kunstdenkmale der Freien und Hansestadt Hamburg*, ed. by GÜNTHER GRUNDMAN, 3 vol. (1953–68).
(He.Th./C.A.McL./Ed.)

Harvey

A leading English physician of the first half of the 17th century, William Harvey achieved fame by his conclusive demonstration of the true nature of the circulation of the blood and the function of the heart as a pump. Functional knowledge of the heart and the circulation had remained almost at a standstill ever since the time of the Greco-Roman physician Galen—1,400 years earlier. Harvey's courage, penetrating intelligence, and precise methods were to set the pattern for research in biology and other sciences for succeeding generations, so that he shares with William Gilbert, investigator of the magnet, the credit for initiating accurate experimental research throughout the world.

By courtesy of Hempstead Parish Church, Essex; photograph, the Royal Academy of Arts, London



Harvey, marble bust by Edward Marshall. In Harvey Chapel, St. Andrews Church, Hempstead, Essex.

William Harvey was born on April 1, 1578, at Folkestone, Kent. His father, Thomas Harvey, was a prosperous businessman and a leading citizen of the small town. William, the eldest of nine children, was the only one to achieve special distinction in his career, but all his brothers were successful in business or at the royal court in London and among them amassed considerable wealth.

Career as physician and scientific innovator. Little is known of William Harvey's boyhood in the Kentish countryside. During the years 1588 to 1593 he was at the King's School attached to the cathedral at Canterbury. In his 16th year Harvey entered Gonville and Caius College, Cambridge, where he was awarded a scholarship in 1593. Although Harvey attended Caius College because of its special interest in educating doctors, his training was grossly inadequate. He was absent from the university for the greater part of his last year (1598–99) because of illness—probably malaria—but had received the B.A. degree in 1597. Determined to continue with medical training, he began a two-and-a-half-year course of study at the University of Padua, reputed to have the best medical school in Europe. His teacher was a celebrated anatomist, Hieronymus Fabricius ab Aquapendente, and it was in the now-famous oval Anatomy Theatre, still to be seen at the university, that Harvey first recognized the problems posed by the function of the beating heart and the properties of the blood passing through it.

Early theories of the heart and circulation

From the time of Aristotle in the 4th century BC it had been widely believed that the blood vessels contained both blood and air. Galen, the Greco-Roman physician, in the 2nd century AD proved that the arteries contained only blood but still believed that air entered the right side of the heart from the lungs. There was a general belief that the movement of the blood was by ebb and flow, an analogy being found in the movement of the sea. Galen's views on this are difficult to assess with exactitude, but it is apparent that he, like everyone else, had no conception of a circular movement of the blood, leaving the heart by one set of vessels, the arteries, and returning to it by another set, the veins. The main propulsive force initiating this oscillatory movement was supposed to be derived from a contracting of the arterial system, rather than by a pumping action of the heart. The blood in the veins was believed to be formed in the liver, passing to the right auricle (*i.e.*, one of the two upper chambers of the heart), and from there to the right ventricle (one of the two lower chambers), to make its way through holes in the septum, or partition, to the left side, where it met with blood from the arteries, which was mixed with air derived from the lungs. This was the extent of man's knowledge about the movement of the blood until 14 centuries later. Early in the 16th century the idea of a pulmonary circulation—that is, a circular motion of blood through heart and lungs—began to occur to some anatomists. In addition, the presence of a perforated septum was beginning to be questioned. In the middle of the 16th century a great anatomist, Andreas Vesalius, also working at Padua, first established accurate knowledge of human anatomy but was less interested in function. Several other medical investigators refined the anatomical knowledge of the heart. Realdus Columbus of Cremona, working as assistant to Vesalius, developed the idea of a pulmonary circulation, and this was made more definite by his pupil, Andreas Caesalpinus, though they still thought that the blood was distributed to the body by the great veins and their branches. Fabricius had a special interest in the anatomy of the veins and first described the system of valves found in them, but he was quite ignorant of their true function. In brief, there existed no convincing explanation of how the heart worked, and Harvey's logical mind remained unsatisfied.

His 28 months at Padua are only meagrely documented, but it is clear that he was outstanding among the students of his year. After receiving his diploma as doctor of medicine of Padua in April 1602, he returned to England. By the standards of the time he was fully trained in anatomy, the simpler functions of the human body, and in therapeutics based on the writings of Aristotle. He had had some clinical experience in the hospitals of Padua and Venice and was entitled to obtain a fellowship of the College of Physicians in London after passing through the preliminary stage of candidate for the higher qualification. At his first oral examination, in May 1603, he was given limited permission to practice medicine, but only after further examinations in April and August 1604 was he fully licensed to practice within the jurisdiction of the college—that is, in the London area.

Shortly after his return to England, Harvey married Elizabeth Browne, daughter of Lancelot Browne, physician to King James I and his queen, and a senior fellow of the college. The couple set up house in the parish of St. Martin's by Ludgate, not far from the College of Physicians; and, backed by Browne, Harvey then tried to obtain the appointment of physician to the Tower of London, where a number of distinguished men were imprisoned. Though he failed in this attempt, in 1607 he finally obtained a fellowship of the College of Physicians, which entitled him to seek an appointment as physician to one of the two great hospitals then serving London—St. Bartholomew's and

St. Thomas's. It may have been through his brother John, who had obtained employment in the King's household, that early in 1609 the King gave Harvey a recommendation for an appointment at St. Bartholomew's, which was conveniently near his house in St. Martin's. He was given the post of assistant physician, and, when the physician died in the summer of that year, Harvey succeeded him.

Appointment to St. Bartholomew's

The hospital at that time had about 200 beds for patients in 12 wards. Harvey's duties consisted of attending in the hall of the hospital to see the patients and prescribe for their treatment; he worked at least one day a week throughout the year and at any other time when specially needed. The physician was usually expected to live within the hospital precincts, but the rule was waived for Harvey since he lived not far away. He received an annual salary of £25 with £2 extra for his livery and a further £8 since he did not use the official residence. His colleagues were three surgeons and an apothecary in charge of the dispensary.

Harvey held this office for 34 years, until 1643 when he was displaced for political reasons by Oliver Cromwell's party, then in power in London. These years saw the development and culmination of his active career as physician and scientific innovator. He developed a large private practice, attending many of the most distinguished citizens, including Sir Francis Bacon—and, about 1618, was made physician extraordinary to King James I, thus becoming a colleague of Sir Theodore Turquet de Mayerne, the senior court doctor. There can be no doubt that Harvey was for many years one of the most widely trusted doctors in England, although his unorthodox views on the circulation of the blood did injure his practice after their publication in 1628. Invariably courteous and regarded with affection and respect by his colleagues, he conducted his practice with common sense and honesty. Though advanced in his ideas of anatomy and physiology and scientific in his methods of research, he was inevitably conservative in the use of remedies. Very few potent drugs were known in his time, and accurate diagnosis was, more often than not, impossible, so that he never escaped from the influence of Aristotle, in whose principles he had been trained. He was the great protagonist of experimental biology but did not apply himself to this form of originality in therapeutics.

At the time of the King's last illness in 1625, de Mayerne was out of the country, and Harvey led the team of doctors in attendance. After the King's death it was rumoured that his favourite, the Duke of Buckingham, had contributed to the fatal outcome by applying remedies not approved by the doctors. He was actually accused of having poisoned the King, and an inquiry was ordered by Parliament in 1626. Harvey was the most important witness of several who contributed to exonerating the Duke from any direct responsibility. Charles I, the new king, continued Harvey's appointment as his personal physician and gave him a special award for the care he had given the last King. Charles's health remained good until the day of his execution, so that he rarely had need to consult his doctor. Nevertheless, Harvey became his close friend and was always in attendance on his journeys, such as his state visits to Scotland in 1633 and 1638. The King helped Harvey's scientific research by putting the deer in the royal parks at his disposal, and he delighted in showing the King anything of curiosity or scientific interest. At the same time, Harvey took his full share in the affairs of the College of Physicians, being constantly present at the meetings of the fellows and occupying all the official positions in the college hierarchy except that of president. His duties at court would not have allowed him to fill this position during his active years, and when it was offered to him in 1654 he was too old and ill to be able to accept.

Yet it is clear from the college records that Harvey was always the man to whom his colleagues turned for advice. The physicians at this time had precedence over the other branches of the profession, and Harvey had a prominent part in maintaining this ascendancy over the surgeons, obstetricians, and apothecaries whenever they became restive under the authority of the college.

Scientific research

In spite of Harvey's activity in medical practice and college affairs, he spent much time in scientific research from the time of his return to England in 1604 until the beginning of the Civil War in 1642. His interest lay primarily in elucidating the facts of the movement of the heart and its relation to the circulation of the blood. Fabricius at Padua had opened his eyes to the value of comparative anatomy, and he was tireless in dissecting every kind of living thing, from insects, earthworms, reptiles, birds, and mammals up to man himself. He seized every opportunity to increase his knowledge of pathology through postmortem examinations and was an acute clinical observer of his patients, not omitting their psychology. Most of his scientific papers were destroyed by parliamentary soldiers during the Civil War, so that there is now no direct evidence of his methods. On the other hand, his lecture notes used from 1616 onward survive. In 1615 he was appointed to a college lectureship intended to cover all parts of medical knowledge, though each lecturer modified the course to suit his own interests. Harvey's manuscript, now in the British Museum, was entitled "Lectures on the Whole of Anatomy." It is written in a very bad hand in mixed Latin and English, and it is incomplete, lacking any account of the skeleton, the sense organs, and other systems. The systematic anatomy is enlivened by many references to comparative anatomy, morbid anatomy, and clinical observations, even naming individuals whom he had treated. It is evident that he wrote these notes before he had come to any conclusions about the circulation of the blood, so that they contain nothing that seriously questioned the authority of Galen. The only reference to his novel views is on a leaf inserted some years later, probably after 1628. Harvey held this lectureship until 1656.

Discovery of circulation of the blood. It is evident from his writings that Harvey reverenced Aristotle, even though he had to dismiss some of his teachings as absurd. He also valued the views of Galen, his predecessor in experimental physiology, and enlisted his support whenever he could do so. Yet Harvey depended essentially on reasoning from his own observations and experiments for proof of his contentions. During the 12 years after 1616, Harvey may have introduced some novelties into his lectures; but, by his own assertions, he demonstrated the results of his researches to his friends privately at the college. In 1628 he finally published his book, *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* (*An Anatomical Exercise Concerning the Motion of the Heart and Blood in Animals*), a slender volume that established the true nature of the circulation of the blood. With Galen's help, he first disposed finally of the idea that the blood vessels contained air. He then elucidated the function of the valves in the heart in maintaining the flow of blood in one direction only when the ventricles contracted—on the right side to the lungs and on the left to the limbs and viscera. He proved that no blood passed through the septum, separating the two ventricles, and he explained the purpose of the valves in the larger veins in directing the return flow toward the heart. He showed that blood was expelled from the ventricles during contraction, or systole, and flowed into them from the auricles during expansion, or diastole. He proved that the arterial pulse was due to passive filling of the arteries by the systole of the heart and not by active contraction of their walls. He explained the purpose of the pulmonary circulation from the right ventricle through the lungs and back to the left auricle and ventricle. His only failure was in not demonstrating the connection of the arterial and venous systems in the tissues of the limbs by means of the smallest, or capillary, vessels. These he was unable to see, having no adequate form of microscope at his disposal. He was the first scientist to employ measurement of the content of the chambers of the heart and estimation of the total amount of blood in the body—that is, quantification.

Fame and criticism

Harvey's book made him famous throughout Europe, though the overthrow of so many time-hallowed beliefs attracted virulent attacks and much abuse from those who did not wish to believe the plain evidence of their senses. He refused to indulge in controversy and made no reply until 1649, when he published a small book answering the criticisms of a French anatomist, Jean Riolan. In this he reiterated some of his former arguments and utterly demolished Riolan's objections.

In 1636 King Charles dispatched a diplomatic embassy to the Holy Roman emperor Ferdinand II at Regensburg, Germany, in an attempt to establish the claim of his nephew, Prince Charles Louis, as Elector Palatine. Harvey was chosen as doctor to the mission and spent ten adventurous months of travel by land and water through territories ravaged by the Thirty Years' War, extending his journey by visits to Vienna, Prague, Venice, Rome, and Naples. At Nuremberg Harvey had a historical encounter with Caspar Hofmann, professor of medicine at the University of Altdorf, whom he attempted, at a public demonstration, to convince of the truth of his doctrine of the circulation. Though he did not succeed, Harvey behaved with great dignity and good temper in the face of obstinate blindness to demonstration of the facts.

At the start of the Civil War in 1642, Harvey was with the King and was in charge of the two princes, Charles and James, in the early stages of the Battle of Edgehill. When the King established his headquarters soon afterward at Oxford, Harvey remained with him and was given the position of warden of Merton College in 1645. Here he resumed his work on the development of the chick in hens' eggs and first met John Aubrey, antiquary and gossip, who afterward left a revealing account of Harvey in his *Brief Lives.* When the defeated King fled from Oxford to surrender himself to the Scots, Harvey joined him for a time at Newcastle but was forced to leave the King when he was handed over to the parliamentary army and was not allowed to go to him when he was imprisoned in the Isle of Wight. Harvey had never been much interested in politics but felt a deep personal regard for the King and after his execution in 1649 was a broken and unhappy man.

Yet two years later he published his second great book. After the publication of *De Motu Cordis,* the main achievement of Harvey's life, he had continued active research into the difficult subject of reproduction in animals. This led in 1651 to the publication of *Exercitationes de Generatione Animalium* (*Anatomical Exercitations Concerning the Generation of Animals*) through the persuasions of his younger friend Sir George Ent, a fellow of the college. The book contains much of historical and scientific interest, but Harvey's thought was greatly influenced by Aristotle. The book is mainly concerned with the development of the chick in hens' eggs, and Harvey insisted throughout that in all living things the origin of the embryo is to be found in the egg. He investigated also the embryology of deer, rejecting Aristotle's notion that menstrual blood played any part in the formation of the fetus; he also questioned whether or not semen had any influence. Having no microscope, he could not see the spermatozoa, which were not demonstrated until 1686 by Leeuwenhoek working in Holland with stronger lenses. Harvey remained uncertain of how fecundation of the ovum was accomplished and even suggested that it was by a kind of infection resembling the origin of infectious diseases. Aristotle had originated the theory of gradual formation of the embryo, part by part, as opposed to the idea of preformation, meaning that all the parts arose in miniature at the same time. Harvey agreed with Aristotle and crystallized the belief in the term epigenesis, though to him its meaning was extremely simple compared with all that is implied by it at the present time. Aristotle believed in the principle of spontaneous generation of primitive organisms; it is probable that Harvey did not support this belief, but his statements are equivocal, and his position remains uncertain.

Harvey's brothers had been successful merchants, and their advice, coupled with his skill as a doctor and his naturally austere habits, enabled him to accumulate considerable wealth. But he had become old and ill. He had met with so much opposition and disbelief that his passionate desire to establish scientific truth was partially unsatisfied. In his last years under Cromwell's Protectorate, he was regarded as a political "delinquent" owing to his long association with King Charles and was forced to spend most of his time lodging in one or another of his brothers' houses outside London. Though he corresponded with many distinguished foreign doctors, he was reluctant to engage in any further scientific research, saw few patients, and took little part in the affairs of the College of Physicians. He showed his regard for the fellows by giving them a new college building in 1652 with a library containing his own collection of books and presumably any remaining manuscripts. This was in use for less than 14 years, being destroyed in the Great Fire of London in 1666, so that very few of his books have survived to the present day. He suffered severe pain from gout and kidney stones and described himself to a correspondent as "not only ripe in years, but also a little weary and entitled to an honourable discharge" from further scientific argument. His last illness was brief. He awoke one morning partially paralyzed and unable to speak, probably owing to a cerebral thrombosis. He died in his 80th year on June 3, 1657, probably in his brother Eliab Harvey's house at Roehampton. He was buried in the family vault at Hempstead, an Essex village 50 miles (80 kilometres) from London. In 1883 he was reburied in a marble sarcophagus in the Harvey Chapel there, near a marble bust by Edward Marshall. This is a lifelike image of Harvey, better, probably, than any of the existing portraits of him in old age.

Last years

BIBLIOGRAPHY. SIR GEOFFREY KEYNES, *The Life of William Harvey* (1966, reissued 1978), is a full and definitive biography based on examination of contemporary sources, documented and illustrated, with eight appendixes; his *A Bibliography of the Writings of Dr. William Harvey, 1578–1657,* 2nd ed. (1953), is an account of all Harvey's books and of where they may be found; and his *The Portraiture of William Harvey* (1949) is a catalog of pictures, genuine and spurious, with reproductions. JOHN G. CURTIS, *Harvey's Views on the Use of the Circulation of the Blood* (1915), is an early study of the position of Harvey's work in the history of the knowledge of human physiology. For texts, see GWENETH WHITTERIDGE (ed.), *The Anatomical Lectures: Prelectiones Anatomie Universalis, De Musculis* (1964), a reliable transcription of Harvey's lecture notes, both in Latin and English, with a full discussion and interpretation, and WHITTERIDGE (trans.), *An Anatomical Disputation Concerning the Movement of the Heart and Blood in Living Creatures* (1976, trans. from Latin); see also WHITTERIDGE, *William Harvey and the Circulation of the Blood* (1970), an important study of the growth of Harvey's ideas. ARTHUR W. MEYER, *An Analysis of the De Generatione Animalium of Harvey* (1936), is a discussion of Harvey's second major publication, a work on animal reproduction and development; *De Generatione Animalium* is also treated in ELIZABETH B. GASKING, *Investigations into Generation, 1651–1828* (1967). WALTER PAGEL, *William Harvey's Biological Ideas: Selected Aspects and Historical Background* (1967), a well-documented historical analysis of Harvey's ideas on physiology and embryology, is continued in his *New Light on William Harvey* (1976). A good survey of Harvey's works is KENNETH D. KEELE, *William Harvey: The Man, the Physician, and the Scientist* (1965). Later studies include JEROME J. BYLEBYL (ed.), *William Harvey and His Age: The Professional and Social Context of the Discovery of the Circulation* (1979); and ROBERT G. FRANK, *Harvey and the Oxford Physiologists: A Study of Scientific Ideas* (1980), an analysis based upon diaries, letters, notebooks, manuscripts, and published scientific works.

(G.L.K./Ed.)

Havana

Havana (Spanish: La Habana) is the capital, major port, and leading commercial centre of Cuba. It also constitutes one of Cuba's 14 provinces. Located on the island's north coast, Havana, with more than 2,000,000 people, is the largest city in the Caribbean region and one of the great treasuries of historic colonial preserves in the Western Hemisphere. Prior to 1959, when Fidel Castro came to power, it was a Mecca for tourists from the United States, who were drawn by the city's many attractions, which included climate and nightlife in addition to history. In the next 30 years, however, despite its continued importance as the island's major economic hub, Havana lost much of its lustre as Castro's socialist government redirected the country's resources primarily toward the improvement of conditions in rural Cuba. Havana's upkeep and improvement thus lagged until the 1980s, when a major rehabilitation project began.

This article is divided into the following sections:

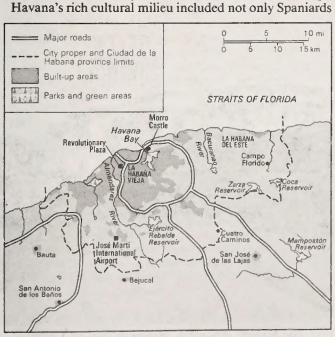
Physical and human geography 479
  The character of the city 479
  The landscape 479
    The city site
    Climate
    The city layout
  The people 480
  The economy 481
    Industry
    Commerce and finance
    Transportation
  Administration and social conditions 481
    Government
    Services
    Health
    Education
  Cultural life 481
History 482
  Foundation and early growth 482
  Development as a major New World port 482
  Alliance with the United States 482
  The city under Castro 482
Bibliography 482

Physical and human geography

THE CHARACTER OF THE CITY

Havana's location along a magnificent deep-sea bay with a sheltered harbour made the city a prime location for economic development from Spanish colonial times in the early 16th century. Cuba is endowed with a number of such harbours, but Havana's on the north coast was prized above the others by the early Spanish colonizers. With land on both sides of the harbour, the port was easily defended. The early colonists erected a number of fortifications in the area that withstood most invaders. In colonial times Havana was the first landfall for Spanish fleets coming to the New World, and it became a staging area, first, for the conquest of the Americas by Spanish conquistadores and, later, for the economic and political domination of the hemisphere by Spain. The city early became a cosmopolitan centre with sprawling fortifications, cobblestone plazas, and buildings with ornamental facades and ornate iron balconies. Today's Havana mixes these structures with a variety of conventional modern buildings.

Havana's rich cultural milieu included not only Spaniards from diverse regions of the Iberian Peninsula but other European peoples as well. The small native Indian population of Cuba was not a significant factor in the Havana area and, in any case, was largely decimated in its early contact with the Spanish. The colonial years brought a large influx of black slaves from Africa who, after the end of slavery in the late 19th century, began flocking to Havana. Today's Havana is a mix of white Spanish stock, black ethnic groups, and significant mulatto strains.



Havana and surrounding area.

THE LANDSCAPE

The city site. The city extends mostly westward and southward from the bay, which is entered through a narrow inlet and which divides into three main harbours: Marimelena, Guasabacoa, and Atarés. The sluggish Almendares River traverses the city from south to north, entering the Straits of Florida a few miles west of the bay. The bay

The low hills on which the city lies rise gently from the deep blue waters of the straits. A noteworthy elevation is the 200-foot- (60-metre-) high limestone ridge that slopes up from the east and culminates in the heights of La Cabaña and El Morro, the sites of colonial fortifications overlooking the bay. Another notable rise is the hill to the west that is occupied by the University of Havana and the Prince's Castle.

Climate. Havana, like much of Cuba, enjoys a pleasant year-round climate that is tempered by the island's position in the belt of the trade winds and by the warm offshore currents. Average temperatures range from 72° F (22° C) in January and February to 82° F (28° C) in August. The temperature seldom drops below 50° F (10° C). Rainfall is heaviest in October and lightest from February through April, averaging 46 inches (1,167 millimetres) annually. Hurricanes occasionally strike the island, but they ordinarily hit the south coast, and damage in Havana is normally less than elsewhere in the country. Hurricanes

The city layout. Walls as well as forts were built to protect the old city, but by the 19th century Havana had already grown beyond the original barriers. The city first spread to the south and west. Expansion to the east was facilitated later by the construction of a tunnel under the entrance to the bay; such suburbs as La Habana del Este were subsequently able to be developed.

Several broad avenues and boulevards stretch across the city. One of the most picturesque is the Malecón, which extends southwestward along the coast from the port en- Major streets

trance to the Almendares River, under which it passes via a tunnel, emerging on the other side in Miramar as Avenida Quinta. Roughly paralleling the Malecón in the Vedado neighbourhood is Linea, another long avenue that passes under the river. Among other thoroughfares of note are Avenida del Puerto, Paseo Martí (or Prado), Avenida Menocal (Infanta), and Avenida Italia.

Old Havana

Contemporary Havana can essentially be described as three cities in one: Old Havana, Vedado, and the newer suburban districts. Old Havana, with its narrow streets and overhanging balconies, is the traditional centre of much of Havana's commerce, industry, and entertainment, as well as being a residential area. It is richly endowed with historic buildings, representing architectural styles from the 16th through the 19th century. Covering some three square miles and hugging the harbour, Old Havana includes Spanish colonial structures, towering Baroque churches, and buildings in Neoclassic style, as well as commercial property and less pretentious homes on the fringes.

To the north and west a newer section, centred on the uptown area known as Vedado, has become the rival of Old Havana for commercial activity and nightlife. This part of the city, built largely in the 20th century, contains attractive homes, tall apartments, and offices along wide, tree-lined boulevards and avenues. It is also the location of many hotels that before 1959 were frequented by U.S. tourists. Central Havana, sometimes described as part of Vedado, is mainly a shopping district that lies between Vedado and Old Havana.

A third Havana is that of the more affluent residential and industrial districts that spread out mostly to the west. Among these is Marianao, one of the newer parts of the city, dating mainly from the 1920s. Some of the suburban exclusivity was lost after the revolution, many of the suburban homes having been expropriated by the Castro government to serve as schools, hospitals, and government offices. Several private country clubs were converted to public recreational centres.

From colonial times Havana has been noted for its parks and plazas. Habaneros, as its residents are called, gather day and night under the sprawling trees of these many green areas. Through colonial times and almost to the end of the 19th century, the Plaza de Armas in Old Havana was the centre of Cuban life. Its most famous building, completed in 1793, is the Palace of the Captains General, an ornate structure that housed the Spanish colonial governors and, from 1902, three Cuban presidents. The building is now a museum.

The Plaza de Armas

In the 1980s many parts of Old Havana, including the Plaza de Armas, became part of a projected 35-year multimillion-dollar restoration project. The government sought to instill in Cubans an appreciation of their past and also to make Havana more enticing to tourists in accordance with the government's effort to boost tourism and thus increase foreign exchange.

One of the first buildings to be restored was the Cathedral of Havana, the church of Havana's patron saint, San Cristóbal (St. Christopher); it was constructed in the 18th century by the Jesuit order. Located near the waterfront, its ornate facade is regarded by art historians as one of the world's finest examples of Italian Baroque design. The restoration work left the cathedral looking much as it did when originally completed.

The Plaza de la Revolución

The expansive Plaza de la Revolución, west of Old Havana, is the site of President Castro's major speeches, which are delivered before crowds of, it is estimated, up to 1,000,000 citizens. The plaza is distinguished by some of the city's most imposing architecture. Surrounding the towering monument to José Martí, leader of Cuban independence, are such modern structures as the National Government Centre, the headquarters of the Communist Party of Cuba and the armed forces, and various government ministries. In Central Havana are more traditional buildings, including the white-domed former National Capitol, now housing the Cuban Academy of Sciences; the Museum of the Revolution, housed in the old Presidential Palace; and the National Museum of Art.

Another restoration project was centred on the old Spanish fortifications that dominate Havana's harbour and, for a time in the 17th and 18th centuries, made Havana the most fortified city in Spanish America. The most famous and impressive of these is Morro Castle (Castillo del Morro), completed in 1640. It became the centre of the network of forts protecting Havana, and, with La Punta Fortress (Castillo de la Punta), dominated the actual entrance to the harbour. The oldest fortification, La Fuerza (Castillo de la Fuerza), was begun in 1565 and completed in 1583. Its site at the Plaza de Armas was that of an even older fort erected by Hernando de Soto in 1538 and later destroyed by French pirates.

Morro Castle

THE PEOPLE

Havana, like the rest of Cuba, is populated mostly by people of Spanish ancestry, with a large minority of blacks and mulattoes, whose ancestors were slaves. There are few mestizos, as in many other Latin-American countries, because the Indian population was virtually wiped out in colonial times. In the era before Fidel Castro came to power, the city was economically and ethnically divided. On the one hand, there was the minority of the wealthy, educated elite, together with a developing and expanding middle class, and on the other was the working-class majority. This division was largely based on ethnic background: whites tended to be more well-to-do, while blacks and mulattoes generally were poor. The economic structure did not provide much opportunity for blacks and mulattoes except in the more menial occupations. There was also little opportunity for them to obtain an education.

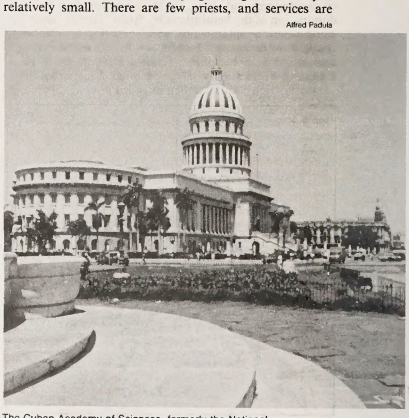
Pre-Castro ethnic divisions

Under the Castro government that came to power in 1959, this system changed. Educational and employment opportunities were made available to Cubans of all ethnic backgrounds. In housing, the government follows an official policy of no discrimination based on ethnic background, and independent observers tend to believe this policy has been more or less faithfully carried out. Where there were few black or mulatto Cubans in middle- to top-level national and local government posts before 1959, there are now many, although still not in the same proportion as in the population. However, many blacks and mulattoes throughout the island are still in a struggle to lift themselves out of poverty.

Habaneros, as Cubans in general, do not constitute a strongly religious community; about half do not profess a religious affiliation. A number of churches in Havana have continued to operate since the Castro revolution. Roman Catholics form the largest religious group, but the number of parishioners worshiping on a given Sunday is relatively small. There are few priests, and services are

Religion

Alfred Padula



The Cuban Academy of Sciences, formerly the National Capitol, in the Central Havana district.

often conducted by lay people. Protestant churches are similarly limited in activity. The Jewish community in Havana is reduced to only a few hundred from once having embraced more than 50,000 people, many of whom had fled Nazi persecution and subsequently left Cuba after Castro came to power.

One of the major phenomena of the Castro era has been the steady flow of Cubans into exile, mainly to the United States but also to Mexico, Venezuela, and elsewhere in Latin America. In the United States the largest concentration of exiles is in Florida, where they have become a significant minority in the state. It is estimated that at least a million Cubans have left the island since 1959, about 60 percent of whom were said to be Habaneros. Heavy emigrations following the revolution created a shortage of professional people in Havana. This was particularly evident in medicine, law, and economics until well into the 1970s. The Cuban government was forced to give high priority to the training of doctors, dentists, lawyers, and economists to replace these professionals.

Emigration after Castro

THE ECONOMY

Havana's economy first developed on the basis of its location, which made it one of the early great trade centres in the New World. Sugar and a flourishing slave trade first brought riches to the city, and later, after independence, it became a renowned resort. Castro, however, de-emphasized the city as a tourist Mecca.

Industry. Despite efforts by the Castro government to spread Cuba's industrial activity to all parts of the island, Havana remains the centre of much of the nation's industry. The traditional sugar industry, upon which the island's economy has been based for three centuries, is centred elsewhere on the island and controls some three-fourths of the export economy. But light manufacturing facilities, meat-packing plants, and chemical and pharmaceutical operations are concentrated in Havana. Other food-processing industries are also important, along with shipbuilding, vehicle manufacturing, production of alcoholic beverages (particularly rum), textiles, and tobacco products, particularly the world-famous Havana cigars. Although the ports of Cienfuegos and Matanzas, in particular, have been developed under the Castro government, Havana remains Cuba's primary port facility; a majority of Cuban imports and exports pass through Havana. The port also supports a considerable fishing industry.

Commerce and finance. Under the Castro government Cuba's traditional free-enterprise system was replaced by a heavily socialized economic system. Most business in Cuba is in the hands of the state. In Havana the large Cuban-owned department stores and U.S.-owned businesses were nationalized and today operate under state control. In Old Havana and throughout Vedado there are a few small private businesses, such as shoe-repair shops or dressmaking facilities, but their number is steadily declining.

Banking has come totally under state control, and the National Bank of Cuba, headquartered in Havana, is the control centre of the Cuban economy. Its branches in some cases occupy buildings that were in pre-Castro times the offices of Cuban or foreign banks.

Transportation. Havana has historically been the hub of Cuba's transportation system and remains so today. This position resulted from Havana's role as the seat of Spanish colonial government and the importance of the port for trade. As railroads developed and as the use of automobiles led to highway construction, the importance of Havana's focal situation increased. Havana became the key terminus for both rail and road links from the east and west. Also, Havana became the main gateway for international air transport. The old Rancho Boyeros airport, now José Martí International Airport, is located eight miles (13 kilometres) from downtown Havana and handles domestic and international flights. A network of bus routes also centres on Havana, and buses are the main mode of inner-city transportation.

Road and railroad development

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Havana is administered by a city council, with a mayor as chief administrative officer. The city is dependent upon the national government, however, for much of its budgetary and overall political direction. The national government is headquartered in Havana and plays an extremely visible role in the city's life. Moreover, the all-embracing authority of many national institutions, including the Communist Party of Cuba (Partido Comunista de Cuba; PCC), the Cuban Army, the militia, and neighbourhood groups called the Committees for the Defense of the Revolution (CDRs), has led to a declining role for the city government, which, nevertheless, still provides such essential services as garbage collection and fire protection. The CDRs, which exist in virtually every street and apartment block, have two main functions: first, to actually defend the revolution against both external and internal opposition and, second, to handle routine tasks in maintaining neighbourhoods.

Role of the CDRs

Havana's borders are contiguous with the province Ciudad de la Habana. Thus Havana functions as both a city and a province. There are two joint councils upon which city and provincial authorities meet—one embraces municipal and provincial leaders on a national basis, the other, a Havana city and provincial council. Havana is divided into 15 constituent municipalities. Until 1976 there were six subdivisions, but in that year the city's borders were expanded to include the entire metropolitan area.

Services. Utility services are under the control of several nationalized state enterprises that have developed since the Castro revolution. Water, electricity, and sewage service are administered in this fashion. Electricity is supplied by generators that are fueled with oil. Much of the original power plant installation, which operated before the Castro government assumed control, has become somewhat outdated. Electrical blackouts occurred, prompting the national government in 1986 to allocate the equivalent of $25,000,000 to modernize the electrical system. It is said that any part of Havana is within five minutes of a fire-fighting unit; the equipment is largely new.

Health. Under the Cuban government all citizens are covered by the national health care plan. Administration of the health care system for the nation is centred largely in Havana. Hospitals in Havana are run by the national government, and citizens are assigned hospitals and clinics to which they may go for attention. During the 1980s Cuba began to attract worldwide attention for its treatment of heart diseases and eye problems, some of this treatment administered in Havana. There has long been a high standard of health care in the city.

Health care quality

Education. The national government assumes all responsibility for education, and there are adequate primary, secondary, and vocational training schools throughout Havana. The government claims that all children receive an education, and the claim appears to be valid. The schools are of varying quality, however, and the content of education is clearly aimed at supporting the socialist political orientation of the Castro government. Education is free and compulsory at all levels except higher learning, which is free nonetheless. The University of Havana, located in the Vedado section of Havana, was established in 1728 and was once regarded as a leading institution of higher learning in the Western Hemisphere. Soon after Castro came to power in 1959, the university lost its traditional autonomy and was placed under the control of the government. The city's only other university, the respected Catholic University in Marianao, was closed after the revolution.

CULTURAL LIFE

Havana, by far the leading cultural centre of the island, offers a wide variety of features that range from museums to ballet and from art and musical festivals to exhibitions of technology. The restoration of Old Havana offered a number of new attractions, including a museum to house relics of the Castro revolution. The government placed special emphasis on cultural activities, many of which are free or involve only a minimal charge.

The Museum of the City of Havana, formerly the Palace of the Captains General in Old Havana, contains many pieces of old furniture, pottery, jewelry, and other examples of colonial workmanship, as well as models of what

Havana looked like in earlier centuries. The museum also houses material relating to the era of U.S. occupation and influence in Cuba. Other important museums are the National Museum of Art in Central Havana and the Museum of Decorative Arts in Vedado. The city's National Library houses Cuba's largest collection. The widest-circulating daily newspapers are published in Havana, but all of these, including the main one, *Granma,* represent Communist Party or government interests. (J.N.G.)

Restaurants and cuisine

Many of the city's finest restaurants are in Old Havana. The most popular is Bodeguita del Medio, once a hangout of Ernest Hemingway. La Floridita, also renowned for its Hemingway associations, claims to be the "birthplace of the daiquiri." In the kitchens of Habanero families, rice, black beans, and bananas are common staples. Although numerous food products are available at special "dollars-only" markets and stores, Habaneros lacking supplemental income (such as tourist dollars or remittances from relatives living abroad) depend almost exclusively on the meagre quotas of food apportioned by the government.

Havana was known in pre-Castro times as the queen city of the Caribbean because of its nightlife and popular culture. Much of that has disappeared, but there is still some nightlife, particularly at the fabled Tropicana nightclub. Its present-day dancers and singers are as gaudily and scantily attired as their predecessors were in pre-Castro times, and the stage settings are big and imaginative.

At Carnival time in July, Cubans express themselves vigorously in dance and song. In Havana, Carnival is now a virtual holiday, with floats and parades officially sanctioned. These floats compete in what has become an annual parade along the waterfront Malecón. Many Cubans are avid sports fans who particularly favour baseball and association football (soccer). Habaneros support a dozen or so baseball teams. The city has several large sports stadiums. Admission to sporting events is generally free, and impromptu games are played in neighbourhoods throughout the city. Social clubs at the beaches provide facilities for water sports and include restaurants and dance halls.

History

FOUNDATION AND EARLY GROWTH

A port called San Cristóbal de la Habana was founded in 1515 by the Spanish conquistador Diego Velázquez de Cuéllar, possibly near the present town of Batabanó on the south coast of the island. It was not a fortuitous choice, for the climate was poor and the region was swampy. Mosquitoes abounded. The site was abandoned in favour of Havana's present location (then called Puerto Carenas) on the north coast in 1519. The natural deepwater port, together with the land protection to the harbour, made Havana a site that early attracted growing numbers of settlers. A royal decree in 1634 recognized its importance, calling it the "Llave del Nuevo Mundo y Antemural de las Indias Occidentales" ("Key to the New World and Rampart of the West Indies"). Havana's coat of arms carries this inscription. The Spaniards began building fortifications, and in 1553 they transferred the governor's residence to Havana from Santiago de Cuba on the eastern end of the island, thus making Havana the de facto capital. English, French, and Dutch sea marauders attacked the city in the 16th century.

The new capital

DEVELOPMENT AS A MAJOR NEW WORLD PORT

During the 17th century eastbound fleets of Spanish ships carrying treasure from the New World rendezvoused at Havana for the trip across the Atlantic to Spain. The port thus became the object of attacks by competing foreign powers and was blockaded several times during that century. By about 1700 the city walls and the major fortifications had been completed. These withstood attacks until, after a three-month siege ending in August 1762, the British under Admiral Sir George Pocock and the Earl of Albermarle took the city as a prize of war. They held it for six months until the treaty ending the Seven Years' War restored Havana to Spain.

That occupation, as onerous as it was to Habaneros, actually stimulated trade between the New World and Europe, and Havana gained new importance as a port, thriving on the sugar and slave trades. Obstacles to commerce by foreigners were gradually removed as the 18th century ended, and all of Cuba, but Havana in particular, began attracting immigrants from countries other than Spain. This, in turn, added new strains to the ethnic mix of the city—French craftsmen, British merchants, German bankers, and others—and gave Havana a distinct international and cosmopolitan character. Wealthier Cuban colonists visited New York City and Philadelphia. A number of U.S. flagships made port calls at Havana, and there was a small but important U.S. trading community established in Havana by 1850. Nevertheless, Cuba remained a Spanish colony despite the wars of independence that raged on the continent in the early 1800s, wars that led to freedom for most of Spain's New World empire.

Growth of U.S. trade

ALLIANCE WITH THE UNITED STATES

Havana, meanwhile, continued to grow as a major world port, rivaling in population and in trade both New York City and Buenos Aires. Cuba became independent from Spain in 1898 with the aid of the United States, and for six decades thereafter Cuba was a close economic and political ally of that country. Increasing numbers of U.S. businesses and tourists were drawn to Havana, which acquired the look of a U.S. city.

Cuba's government through these years was frequently a dictatorship; at best, it was a fragile democracy, with corruption often running rampant. Many Cubans were unhappy both with the U.S. influence and with the continued dictatorship. There were a number of abortive coup attempts against the government of Fulgencio Batista, but change did not come until the revolution led by Fidel Castro, who on Jan. 1, 1959, took control of Cuba.

Castro takes power

THE CITY UNDER CASTRO

Havana's close ties with the United States were quickly ended, and Castro turned to the Soviet Union for economic and military assistance. Soviet vessels frequented the port of Havana, Soviet-made automobiles and trucks became common in the streets, and Soviet or Soviet-bloc goods appeared in stores.

Despite these influences, Havana still retains many of its older traditions, particularly in the narrow streets of Old Havana where past and present merge. This older part of the city deteriorated somewhat after the revolution as Castro first directed resources to the hinterlands. However, in the early 1980s Old Havana was designated a UNESCO World Heritage site, and the city began a revival as several historic buildings were restored.

Havana's economy was nearly crippled following the dissolution of the Soviet Union in 1991. Trade with the Soviet Bloc had effectively subsidized the Cuban economy, and the loss of this income translated into shortages of goods, power, and transportation. The government responded to the crisis by relaxing controls on private enterprise (including family restaurants and farmers' markets) and permitting U.S. dollars to circulate. It also promoted foreign investment, notably in the tourist sector, resulting in markedly increased tourist traffic by the early 21st century. Despite these changes, Habaneros remain largely focused on their families and neighbourhoods, numerous educational and cultural events are still promoted, and the government retains direct or indirect influence at nearly all levels of society, from home ownership to medical care. (J.N.G./Ed.)

BIBLIOGRAPHY. Havana's role in Cuban national life is discussed in SUSAN EVA ECKSTEIN, *Back from the Future: Cuba Under Castro* (1994). Travel books include GÜNTER GRAU, *Havana,* trans. from German (1985); and *Havana & the Best of Cuba* (2001). Daily life, economic and societal changes, and architectural restoration are examined in two articles in *National Geographic* (June 1999) by JOHN J. PUTMAN, "Cuba," pp. 2–35; and A.R. WILLIAMS, "The Rebirth of Old Havana," pp. 36–45. Historical topics are treated in JOHN ROBERT MCNEILL, *Atlantic Empires of France and Spain: Louisbourg and Havana, 1700–1763* (1985); DAVID SYRETT (compiler), *The Siege and Capture of Havana, 1762* (1970); and MARÍA LOBO MONTALVO, ZOILA LAPIQUE BECALI, and ALICIA GARCÍA SANTANA, *Havana: History and Architecture of a Romantic City* (2000). (Ed.)

Hebrew Literature

Literature in Hebrew has been produced uninterruptedly from the early 12th century BC, and certain excavated tablets may indicate a literature of even greater antiquity. From 1200 BC to *c.* AD 200, Hebrew was a spoken language in Palestine, first as biblical Hebrew, then as Mishnaic Hebrew, a later dialect that does not derive directly from the biblical dialect and one that gained literary status as the Pharisees began to employ it in their teaching in the 2nd century BC. It was not revived as a spoken language until the late 19th century, and in the 20th century it was adopted as the official language of the new State of Israel. The latter event gave impetus to a growing movement in Hebrew literature centred in Israel.

Hebrew literature is not synonymous with Jewish literature. Some Hebrew writing was produced by the Samaritans and in the 17th century by Protestant enthusiasts. Jews also produced important literatures in Greek, Aramaic, Arabic, Judeo-Spanish (Ladino), Yiddish, and a number of other languages. Apart from the Aramaic writings, however, such literatures always served only that part of Jewry that used the language in question. When that community ceased to exist, the literature produced in languages other than Hebrew was forgotten (or, in the case of Greek Jewish literature, became part of Christian tradition) except for whatever part of it had been translated into Hebrew and thus became part of Hebrew literature.

The Hebrew language, though not spoken between *c.* AD 200 and the late 19th century, has always adapted itself to the needs of changing literary tastes. In the Bible it develops from a simple and earthy idiom to a language suitable for the expression of sophisticated religious thought without losing the poetic force and rhythmic fullness that characterizes it. Mishnaic Hebrew is pedestrian, exact, and yet can reach heights of irony or of warmth. In medieval poetry Hebrew allows extravagant displays of verbal artistry but also, in northwestern Europe, a simplicity equal to that of the spoken languages of its milieu. One generation of translators in the 12th century created a scientific Hebrew that is not inferior to contemporary Arabic or Latin in precision or syntactic refinement. The 17th–19th centuries saw the formation of a stately, rigid, classical style based on biblical Hebrew, but at the same time eastern European mystics made the language serve the expression of their love of God. Literary Hebrew in the 20th century draws upon ancient literature to a marked degree, with styles often modeled upon ancient predecessors. The modern period has also evolved a new type of language for nonliterary writing, while in novels the style is often based upon the spoken language.

This article is divided into the following sections:

Ancient Hebrew literature 483
  Preexilian period, *c.* 1200–587 BC
  Period of the Second Temple, 538 BC–AD 70
  Talmudic literature
Literary revival, 500–1000 484
  Piyyuṭim
  Adoption of Arabic metre
The Middle Ages 484
  The Palestinian tradition in Europe, 800–1300
  The golden age in Spain, 900–1200
The period of retrenchment, 1200–1750 484
  Hebrew culture in western Europe
  Eastern Europe and the religious crisis
The 18th and 19th centuries 485
  Beginnings of the Haskala movement
  Romanticism
Modern literature in Hebrew 485
  Formative influences
  Émigré and Palestinian literature
  Israeli literature
Bibliography 486

ANCIENT HEBREW LITERATURE

Preexilian period, c. 1200–587 BC. All that is preserved of the literature of this period is slightly more than 20 of the 39 books included in the Old Testament (the remainder being from the next period). Poetry probably preceded prose. Biblical poetry was based on the principle of parallelism; *i.e.*, the two halves of a verse express the same idea, either by repeating it in different words or by stressing different aspects of it. Examples are found in the book of Psalms: "But they flattered him with their mouths; they lied to him with their tongues" (Ps. 78:36); "He turned their rivers to blood, so that they could not drink of their streams" (Ps. 78:44). To this form was added a simple rhythm, consisting mainly in having each half of a line divided into an equal number of stressed words. There were also folk songs, to which belonged perhaps large parts of the Song of Solomon, dirges, epic chants, and psalms. The use of various forms of poetry in the work of the prophets appears to be a later development.

The principle of parallelism in biblical poetry

The earlier prose texts were still very close to poetry in structure and language. The first real prose may well have been some of the laws recorded in the Pentateuch. In Jeremiah and Deuteronomy a high standard of prose rhetoric was achieved: some of the conversations in the historical books were attempts to reproduce in writing the style of ordinary speech. (See also BIBLICAL LITERATURE.)

Period of the Second Temple, 538 BC–AD 70. The literary output of this period was large, only part of it belonging to the biblical canon. The biblical Hebrew of the writings was artificial because it had ceased to be spoken and had been replaced by Aramaic, a related Semitic language, and Mishnaic Hebrew. Works that are included among the Dead Sea Scrolls belong to this period. Some of these works provide evidence of a new kind of writing, the homiletic, or sermonizing, commentary to the Bible called Midrash. The only work of real literary merit among the scrolls is the fervent personal poetry of the Hymns of Thanksgiving.

Parts of the biblical books of Ezra and Daniel and certain works among the Dead Sea Scrolls are in an early form of Aramaic. This period also began to provide translations (called Targums) of most of the Hebrew Bible into a slightly later Aramaic.

Talmudic literature. In contrast to the works of the Bible and the Second Temple were the collections of writings concerned with Jewish civil and religious law. Whereas the former were lengthy writings bearing the imprint of their authors or editors, early rabbinic literature consisted entirely of collections of individual statements loosely strung together. The individual paragraphs exhibit the influence of Hellenistic rhetoric. Collections that follow the arrangement of biblical books are called Midrash, as opposed to works such as the Mishna, where the material is arranged according to subject. The Mishna was the main work of the period *c.* 100 BC–AD 200. The following period, AD 200–500, was notable for two main innovations: the appearance of an additional literary centre in Babylonia, where Jewry flourished in contrast to its subjugation under the oppressive rule of Rome and, later, Byzantium in Palestine; and the literary use of the spoken local dialects of Aramaic alongside Hebrew. The Talmuds produced by Palestine and Babylonia in this period contained a large

The Mishna and Midrash

proportion of Haggada, statements dealing with theological and ethical matters and using stories, anecdotes, and parables to illustrate certain points. This material was later an influence on Hebrew fiction of the Middle Ages and of the modern period. (See also JUDAISM: *The literature of Judaism.*)

LITERARY REVIVAL, 500–1000

In the 6th century, some Jewish groups attempted to enforce the exclusive use of Hebrew in the synagogue, this tendency being part of a Hebrew revival that began in Palestine and spread westward but did not reach Babylonia until the 10th century.

Piyyuṭim. Synagogues began in this period to appoint official precentors, part of whose duty it was to compose poetical additions to the liturgy on special sabbaths and festivals. The authors were called *paytanim* (from Greek *poiētēs*, "poet"), their poems *piyyuṭim*. The keynote was messianic fervour and religious exuberance. Besides employing the entire biblical, Mishnaic, and Aramaic vocabularies, the *paytanim* coined thousands of new words. Such poems, presupposing a highly educated audience, abound in recondite allusions and contain exhaustive lists of rites and laws. It is known that the most outstanding poets—Phineas the Priest, Yose ben Yose, Yannai, and Eleazar ha-Kalir, or ben Kalir—lived in that order, but when or where in Palestine any of them lived is not known. The accepted datings are 3rd century and 5th–6th century AD. Many *piyyuṭim* are still used in the synagogue.

Adoption of Arabic metre. Biblical Hebrew was re-established as the literary idiom about 900 by Sa'adia ben Joseph, grammarian and religious polemicist. The Arabic system of quantitative metre was adapted for Hebrew during this period (900–1000), probably by Dunash ben Labrat. At first the *piyyuṭ* form was retained for religious poems, and the new metres were used only for secular poetry, which closely imitated Arabic models and, like the latter, was chiefly employed for laudatory addresses to prominent people.

THE MIDDLE AGES

The Palestinian tradition in Europe, 800–1300. From Palestine, the Hebrew renaissance soon spread into the Byzantine Empire. In Sicily and southern Italy (which belonged to Byzantium) several important *paytanim* were at work, and before 1000 a secular literature began to arise in Italy: a fantastic travelogue of Eldad the Danite; a historical romance, *Sefer ha-yashar* (1625; Eng. trans., *Sefer ha-yashar, the Book of the Righteous*) and *Josippon*, a revision of Josephus' *Antiquities* filled with legendary incidents—this last-named book was popular until modern times and was translated into many languages. Nathan ben Yehiel completed in 1101 at Rome a dictionary of Talmudic Aramaic and Hebrew, the *'Arukh*, which is still used.

In the middle of the 10th century members of the north Italian family Kalonymos brought Talmudic studies and *piyyuṭim* to Mainz, Ger., where the yeshiva (school) became a centre of studies under the direction of Gershom ben Judah, known as "the Light of the Exile." As a poet, he established a distinctive style of European *piyyuṭ* in poems that read very much like early European popular poetry. The greatest alumnus of the Mainz academy was Rashi, an author of complete commentaries on the Bible and on the Babylonian Talmud, himself a poet of note.

The slaughter of Jewish peoples in western and central Europe during the Crusades drove large masses of Jews into eastern Europe. The German Jews carried with them their Yiddish speech but hardly any literary culture. In Germany accounts of the disaster were written in a new prose style permeated with poetry; liturgical poetry became henceforth mainly a chronicle of persecutions. These sufferings inspired an important mystical movement, largely propagated through stories, of which the chief collections are the *Ayn Shoyn Mayse Buch* (1602; *Ma'aseh Book*) and the *Sefer Ḥasidim* (1538; "The Book of the Just"), the latter attributed to Judah ben Samuel, "the Hasid" of Regensburg (died 1217).

The golden age in Spain, 900–1200. Spanish Jewry began to flourish in Muslim Spain under the caliphate of Córdoba, where Hasdai ibn Shaprut, a vizier, was the first great patron of Hebrew letters. His secretary, Menahem ben Saruk (died *c.* 970), wrote a biblical lexicon, which was criticized by Dunash ben Labrat when the latter arrived in Spain with philological ideas from the East. Samuel ha-Nagid, vizier of Granada (990–1055), himself a poet and philologist, gathered around him a group of poets, most outstanding among whom was Ibn Gabirol. Moses ibn Ezra of Granada (died *c.* 1139) was the centre of a brilliant circle of poets. Moses' kinsman Abraham ibn Ezra, a poet, philosopher, grammarian, and Bible commentator, attacked the language and style of the early *paytanim;* he and Judah ben Samuel Halevi were the first to use Arabic metres in religious poems. Dominated by Arab standards of taste, the secular poetry dealt with themes of Arabic poetry and often reproduced Arabic phrases; it was written to be appreciated by a small circle of connoisseurs and declined with the collapse of Jewish prosperity in Muslim Spain. The last major poet in Spain was Judah ben Solomon Harizi, who translated various philosophical works into Hebrew. The poets of Granada

The use of biblical Hebrew was made possible by the work of philologists. Of great importance was the creation of comparative linguistics by Judah ibn Kuraish (about 900) and Isaac ibn Barun (about 1100). Judah Hayyuj, a disciple of Menahem ben Saruk, recast Hebrew grammar, and, in the form given to it by David Kimhi of Narbonne (died *c.* 1235), the new system was taken over by the Christian humanists and through them by modern scholarship. The first complete Hebrew grammar, *Kitāb al-luma'* (1886; "The Book of the Variegated Flower Beds"), was written by Ibn Janāḥ of Córdoba (died 1050).

Jewish medieval philosophers in Spain wrote in Arabic, not Hebrew, until the 13th century. Apart from Isaac Israeli (north Africa, died *c.* 940) few medieval Jews made original contributions to science, but the Spanish Jews shared the best scientific education. Abraham bar Hiyya (died *c.* 1136) of Barcelona was an original mathematician who wrote in Hebrew works on mathematics, astronomy, and philosophy. When the Almohads expelled the Jews from Muslim Spain in 1148, many learned refugees went to Languedoc and Provence and there translated scientific and philosophical works.

THE PERIOD OF RETRENCHMENT, 1200–1750

Hebrew culture in western Europe. From 1200 to 1750 was the era of the ghetto, during which the area of western European Hebrew culture shrank to a remnant in Italy, while an entirely different culture arose in eastern Europe. The appearance in 1200 of the Hebrew version, translated from Arabic, of Moses Maimonides' *Moreh Nevukhim* (1851–85; *The Guide of the Perplexed*), which applied Neoplatonic and Aristotelian philosophy to biblical and rabbinic theology, provoked orthodox circles into opposition to all secular studies. As a result of Maimonides' work, there was a return to Neoplatonist mysticism in a form known as Kabbala. This culminated in the theosophy of the *Zohar* (1560; "The Book of Splendor"), which is ascribed to Moses de Leon and which exercised an influence comparable only with that of the Bible and Talmud. Hebrew culture, however, was reduced to a miniature scale in the West after the expulsion of the Jews from England (1290), from France (1306), and from Spain (1492). It continued in Italy, where it remained in contact with contemporary Christian thought. The most outstanding figure was the mystical philosopher Moshe Ḥayyim Luzzatto, who wrote a work on poetics and three remarkably modern plays. Kabbala

Eastern Europe and the religious crisis. In the kingdom of Poland (which then extended from Lithuania to the Black Sea) refugees from German persecution mingled with earlier Byzantine émigrés to create, by the 15th century, a prosperous Jewry with extensive autonomy. Their culture was not a continuation of western European Hebrew civilization but a new creation. The Bible (except for the Pentateuch) was neglected, while the Babylonian Talmud—hitherto studied only by specialists—became the basis of all intellectual life, particularly since the so-called

pilpul method of Jacob Pollak had turned its study into an exciting form of mental gymnastics. The typical literature consisted of *novellae* (*hiddushim*), ingenious discussions of Talmudic minutiae written in an ungrammatical mixture of Hebrew and Aramaic. Imaginative literature existed only in Yiddish, for women and the uneducated.

The expulsion from Spain produced a wave of messianic emotion. Kabbala flourished in Safad, the new Palestinian centre, the meeting place of Spanish, European, and Oriental Jews. There, in 1570–72, Isaac Luria created a cosmic messianism. Though its formulation, in the writings of his pupil Ḥayyim Vital, was abstruse and esoteric, its phraseology penetrated the widest masses, as a result of the introduction of Kabbalist prayers, and coloured all later Hebrew writing. Luria's teachings were developed by the false messiah Sabbatai Zebi in the next century, for and against whom a vast literature was written.

The sufferings of Polish Jewry in the Cossack massacres of 1648—described in a long poem by the Talmudist Yom Ṭov Lipmann Heller—opened their country to Lurianic mysticism. Out of popular Kabbalist elements, Israel ben Eliezer, called the Ba'al Shem Ṭov, produced Ḥasidism. His teaching, like that of his successors, was oral and, of course, in Yiddish; but it was noted by disciples in a simple, colloquially flavoured Hebrew. Since they taught mainly through parables, this may be considered to mark the beginning of the Hebrew short story. Indeed these narratives exercised, and still exercise, a profound influence on modern Hebrew writers.

Ḥasidism

THE 18TH AND 19TH CENTURIES

In the 18th century the conservative mystical movement of Ḥasidism spread rapidly over all eastern Europe except Lithuania. There, Elijah ben Solomon of Vilna, a writer of unusually wide scope, advocated a better graded course of Talmudic training. Shneur Zalman of Ladi created the highly systematized Ḥabad Ḥasidism, which was widely accepted in Lithuania. The Musar movement of Israel Salanter encouraged the study of medieval ethical writers.

Beginnings of the Haskala movement. In the Berlin of Frederick II the Great, young intellectuals from Poland and elsewhere, brought in as teachers, met representatives of the European Enlightenment; they came under the influence of Moses Mendelssohn and also met some representatives of Italian and Dutch Hebrew cultures. One, a Dane, Naphtali Herz Wessely, who had spent some time in Amsterdam, wrote works on the Hebrew language, and another, an Italian, Samuel Aaron Romanelli, wrote and translated plays. Out of these contacts grew Haskala ("Enlightenment"), a tendency toward westernization that venerated Hebrew and medieval western Jewish literature. Among German Jews, then already in rapid process of Germanization, this Hebrew movement had no place. The Enlightenment was introduced in Galicia (Austrian Poland), a centre of Ḥasidism, by the Edict of Toleration (1781) of the emperor Joseph II. By supporting some of its aims, Hebrew writers incurred hatred and persecution. Their chief weapon was satire, and the imitation by Joseph Perl of the *Epistolae obscurorum virorum* (1515; "Letters of Obscure Men") of Crotus Rubianus and the essays of Isaac Erter were classics of the genre. One poet, Meir Letteris, and one dramatist, Naḥman Isaac Fischman, wrote biblical plays.

Romanticism. Galicia's chief contribution was to the Jüdische Wissenschaft, a school of historical research with Romanticist leanings. The impact of Haskala ideas upon the humanistic Italo-Hebrew tradition produced a short literary renaissance. Its main connections were with the Jüdische Wissenschaft, to which Isaac Samuel Reggio contributed. Samuel David Luzzatto, a prolific essayist, philologist, poet, and letter writer, became prominent by his philosophy of Judaism, while a poet, Rachel Morpurgo, struck some remarkably modern chords. For the Jews of the Russian Empire, the Enlightenment proper began with Isaac Baer Levinsohn in the Ukraine and with Mordecai Aaron Ginzberg (Günzburg), in Lithuania. In the 1820s an orthodox reaction set in, coinciding with the rise of a Romanticist Hebrew school of writers. A.D. Lebensohn wrote fervent love songs to the Hebrew language, and his son Micah Joseph (Mikhal), the most gifted poet of the Haskala period, wrote biblical romances and pantheistic nature lyrics. The first Hebrew novel, *Ahavat Ziyyon* (1853; "The Love of Zion"), by Abraham Mapu, was a Romantic idyll, in which Mapu, like all Haskala writers, employed phrases culled from the Bible and adapted to the thought the writer wished to express.

First Hebrew novel

Mapu's third novel, *'Ayit tzavua'* (1857–69; "The Hypocrite"), marked a departure. It dealt with contemporary life and attacked its social evils and portrayed a new type, the *maskil* (possessor of Haskala), in a fight against orthodox obscurantism. The new, aggressive Haskala soon came under the influence of Russian left-wing writers, such as Nikolay Gavrilovich Chernyshevsky and Dmitry Pisarev. Judah Leib Gordon, like Mapu, had started as a Romantic writer on biblical subjects. From 1871 onward he produced a series of ballads exposing the injustices of traditional Jewish life. Moses Leib Lilienblum began as a moderate religious reformer but later became absorbed by social problems, and in *Mishnat Elisha ben Abuyah* (1878; "The Opinions of Elisha ben Abuyah") he preached Jewish socialism. Peretz Smolenskin created in six novels a kaleidoscope of Jewish life in which he rejected the westernized Jew as much as orthodox reactionaries did.

MODERN LITERATURE IN HEBREW

Formative influences. The first formative influences on 20th-century Hebrew literature belong to the late 19th century. The middle classes of eastern European Jewry that read Hebrew books turned to Jewish nationalism, and Zionist activity, coupled with the movement for speaking Hebrew, widened the circle of Hebrew readers. Hebrew daily papers began to appear in 1886. Writers borrowed extensively from medieval translators and European languages, and the Hebrew language assumed a new character. A key figure in the transition to modern writing was Shalom Jacob Abramowitsch, who wrote under the pseudonym of Mendele Mokher Sefarim; after his first novel he became convinced that biblical Hebrew was unsuitable for modern subjects and turned to Yiddish. From 1886 onward he returned to writing mainly in Hebrew and by using Hebrew and Aramaic phrases from the Talmud was able to capture the homeliness he prized in Yiddish. His stories depicted life as it really was, and his style and support of traditional values attracted a wide readership. The popularity of his stories of ghetto life ensured that they would remain the most read and written genre of Hebrew literature until the mid-20th century. A group of writers adopted "grandfather Mendele" as their model. One of these, Asher Ginzberg (Aḥad Ha'am), wrote, from 1889 onward, articles evolving a secular philosophy of Jewish nationalism. His periodical *ha-Shiloaḥ* attained editorial standards previously unknown in Hebrew. From 1921, he devoted his last years to the editing of his correspondence, a valuable documentary of the period.

Depiction of ghetto life and the trend toward nationalism

Ḥayyim Naḥman Bialik, an important poet, essayist, editor, and anthologist of medieval literature, was, for a time, literary editor of *ha-Shiloaḥ* and was much influenced by Aḥad Ha'am. His poetry expressed the inner struggles of a generation concerned about its attitude to Jewish tradition. Saul Tchernichowsky, on the other hand, was untroubled by tradition, and his poetry dealt with love, beauty, and the three places where he had lived: the Crimea, Germany, and Palestine. Isaac Leib Peretz, who wrote both in Hebrew and in Yiddish, introduced the Ḥasidic, or pietistic devotional, element into literature. The emotionalism and simple joy of life of that milieu thereafter strongly influenced writers, and the language absorbed many Ḥasidic terms. A literary historian, Ruben Brainin, discerned the presence of a "new trend" in literature and foresaw a concentration on human problems. Bialik had already pointed to a conflict between Judaism and the natural instincts of Jews. This psychological interest dominated the work of a group of short-story writers and, in particular, that of the writer and critic David Frischmann, who, more than anyone else, imposed European standards on Hebrew literature. European literary tendencies thus became absorbed into Hebrew. Uprooted by the pogroms of 1881 and the two Russian revolutions of 1905 and 1917,

The conflict between Judaism and the natural instincts of the Jews

Jews had emigrated to western Europe and America, and Hebrew literary activity in eastern Europe was disrupted. The Soviet Union eventually banned Hebrew culture, and it also decayed in other eastern European countries and in Germany as the position of Jews deteriorated.

Émigré and Palestinian literature. The writers of this generation were known as the émigré writers. Their work was pessimistic, as the rootlessness without hope of Uri Nissan Gnessin and Joseph Ḥayyim Brenner exemplified. The majority of writers active in Palestine before 1939 were born in the Diaspora (Jewish communities outside Palestine) and were concerned with the past. An exception was Yehuda Burla, an Oriental Jew who wrote about Oriental Jewry. The transition from ghetto to Palestine was achieved by few writers, among them Asher Barash, who described the early struggles of Palestinian Jewry. Shmuel Yosef Agnon, the outstanding prose writer of this generation (and joint winner of the 1966 Nobel Prize for Literature), developed an original style that borrowed from the Midrash (homiletical commentaries on the Hebrew Scriptures), stories, and ethical writings of earlier centuries. Whereas his earlier stories were set in Galicia, he began in the 1940s to write about Palestine.

In contrast, poetry immediately addressed Palestinian life. Among outstanding writers were Rachel (Rachel Bluwstein), who wrote intensely personal poems; Uri Zevi Greenberg, a political poet and exponent of free verse; and Abraham Shlonsky, who led the Symbolist school.

Israeli literature. World War II and the Arab–Israeli War of 1948–49 brought to the fore Palestinian-born writers who dealt with the problems of their generation in colloquially flavoured Hebrew. In the State of Israel, where Hebrew had become the official language, literature developed on a large scale, mainly along contemporary western European and American lines. The extreme diversity in culture of parts of the population, and the problems of new immigrants, provided the main theme for fiction. Poetry flourished, but original drama at first was slow to develop. Greenberg's *Rehovot HaNahar* (1951; "Streets of the River") traced the process by which the humiliation of the massacred is transmuted by the pride of martyrdom into the historical impulse of messianic redemption. In a long dramatic poem, "Bein ha-Esh ve-ha-Yesha" (1957; *Between the Fire and Salvation*), Aaron Zeitlin envisioned the annihilation of European Jewry in mystical terms, examining the relationship of catastrophe and redemption.

Native Israeli prose writers wrote of their life in the kibbutz, the underground, and the war of 1948–49. S. Yizhar and Moshe Shamir emerged as the outstanding representatives of this generation, probing the sensibility of the individual in a group-oriented society. But the establishment of the State of Israel could not allay the anxieties of the individual. The dominant themes of writers who had no access to collective ideals were personal ones—frustration, confusion, and alienation. The poetry of this period and of following decades, represented, among others, by Yehuda Amichai and Haim Gouri, also emphasizes the dissolution of social coherence and expresses the individual devoid of a sense of historical and spiritual mission. The novelist Aharon Megged's *ha-Hai 'al ha-met* (1965; *The Living on the Dead*) casts a putative hero of the pioneer generation in an ironic light.

Memories of the Holocaust haunt the lyrical work of Aharon Appelfeld. Flight and hiding are the characteristic situations of his early stories. His *Badenhaim, 'ir nofesh* (*Badenheim 1939*), published in 1975, captures the ominous atmosphere of the approaching Holocaust sensed by a group of assimilated Jews vacationing at an Austrian resort. It describes social and spiritual disintegration as does his novel *Tor ha-peli 'ot* (1978; *The Age of Wonders*). In the stories of Abraham B. Yehoshua the narrator's tone is remote, the people are drained of emotion. Occasionally an act of feeling or meaning breaks the mood of boredom and illuminates a character's humanity. In both *ha-Mé'ahev* (1976; *The Lover*) and *Gerushim me'uharim* (1982; *A Late Divorce*) Yehoshua explores the confrontation between the philosophy of the present generation and the ideology of the Zionist founders. Personal frustration and religious vision are the subjects of the novelist Pinḥas Sadeh. Yitzḥak Orpaz's novels tend toward psychological exploration, particularly in the series beginning with *Bayit le-adam eḥad* (1975; "One Man's House"). Yoram Kaniuk examines the alienated Israeli but, in *ha-Yehudi haaḥaron* (1981; *The Last Jew*), explores the Israeli experience as a response to the Holocaust. Yitzḥak Ben Ner sets realistic stories in rural and urban communities (*Shekí'ah kefarit* [1976; "A Rustic Sunset"] and *Erez reḥokah* [1981; "A Distant Land"]). Ya'akov Shabtai's novel *Zikhron devarim* (1977; *Past Continuous*) broke new ground in its evocation of the family in society. The world of Amos Oz is generally symbolized by a kibbutz besieged by a powerful and primitive external force always threatening to break in. His *Menuḥah nekhonah* (1982; *A Perfect Peace*) considers the notion of escape from ideological and geographical confinement. Amalia Kahana-Carmon explores the subjective impressions of experience and the complexities of time and memory through a stream-of-consciousness technique.

BIBLIOGRAPHY. JULIUS A. BEWER, *The Literature of the Old Testament*, 3rd ed., completely rev. by EMIL G. KRAELING (1962); SANFORD CALVIN YODER, *Poetry of the Old Testament* (1948, reprinted 1973), a useful introduction and representative collection of biblical poetry; MEYER WAXMAN, *A History of Jewish Literature*, 2nd ed., 5 vol. (1938–60), excellent coverage of Yiddish writing as well as Hebrew; ABRAHAM E. MILLGRAM (ed.), *An Anthology of Medieval Hebrew Literature* (1961); BENZION HALPER, *Post-Biblical Hebrew Literature: An Anthology*, 2 vol. (1921, reprinted 1946); LEON I. FEUER, *Jewish Literature Since the Bible* (1937), with supplements by AZRIEL EISENBERG; T. CARMI (ed.), *The Penguin Book of Hebrew Verse* (1981), a representation of Hebrew poetry from the Bible to the present day, with a useful introduction; MENACHEM RIBALOW, *The Flowering of Modern Hebrew Literature*, trans. from the Hebrew and ed. by JUDAH NADICH (1959); JOSEPH KLAUSNER, *A History of Modern Hebrew Literature, 1788–1930* (1932, reprinted 1972; originally published in Hebrew, 1920); JACOB S. RAISIN, *The Haskalah Movement in Russia* (1913, reprinted 1972); DOV VARDY, *The New Hebrew Poetry* (1947); SAMUELE AVISAR, *Teatro Ebraico* (1957); DAVID PATTERSON, *The Hebrew Novel in Czarist Russia* (1964); MENACHEM RIBALOW, *Anthology of Hebrew Poetry in America* (1938), with text in Hebrew; SIMON HALKIN, *Modern Hebrew Literature, from the Enlightenment to the Birth of the State of Israel: Trends and Values*, new ed. (1970); JACOB J. PETUCHOWSKI, *Theology and Poetry: Studies in the Medieval Piyyut* (1978), with a general introduction, text, and notes; STANLEY BURNSHAW, T. CARMI, and EZRA SPICEHANDLER (eds.), *The Modern Hebrew Poem Itself* (1965), in Hebrew with English transcription and rendering; EISIG SILBERSCHLAG, *From Renaissance to Renaissance*, vol. 2: *Hebrew Literature in the Land of Israel, 1870–1970* (1977); GLENDA ABRAMSON, *Modern Hebrew Drama* (1979); ELLIOTT ANDERSON (ed.), *Contemporary Israeli Literature: An Anthology* (1977); JACOB SONNTAG (ed.), *New Writing from Israel 1976: Stories, Poems, Essays* (1976).

(Ch.R./S.Lr./G.M.A.)

Hegel and Hegelianism

G.W.F. Hegel was the last of the great philosophical-system builders of modern times. His work, following upon that of Immanuel Kant, Johann Gottlieb Fichte, and Friedrich Schelling, thus marks the pinnacle of classical German philosophy. As an absolute Idealist inspired by Christian insights and grounded in his mastery of a fantastic fund of concrete knowledge, Hegel found a place for everything—logical, natural, human, and divine—in a dialectical scheme that repeatedly swung from thesis to antithesis and back again to a higher and richer synthesis. His influence has been as fertile in the reactions that he precipitated—in Søren Kierkegaard, the Danish Existentialist; in the Marxists, who turned to social action; in the Vienna Positivists; and in G.E. Moore, a pioneering figure in British Analytic philosophy—as in his positive impact.

This article deals with the man and his accomplishments as well as with the philosophical movement, Hegelianism, that evolved from his thought. It is divided into the following sections:

Life and work 487
- Early life 487
  - Emancipation from Kantianism
  - Career as lecturer at Jena
- Gymnasium rector 489
- University professor 489
  - At Heidelberg
  - At Berlin
- Personage and influence 490

Hegelianism 491
- General considerations 491
  - Problems of the Hegelian heritage
  - Stages in the history of the interpretation of Hegel
- Crises in the earlier Hegelian school 491
  - Polemics during the life of Hegel: 1816–31
  - Period of controversies chiefly in religion: 1831–39
  - Period of atheistic and political radicalism: 1840–44
- Hegelianism through the 20th century 494
  - Development and diffusion of Hegelianism in the later 19th century
  - Hegelianism in the first half of the 20th century
  - Hegelian studies today

Bibliography 496

Life and work

EARLY LIFE

Hegel—who was born in Stuttgart on August 27, 1770, the son of a revenue officer—was christened Georg Wilhelm Friedrich. He had already learned the elements of Latin from his mother by the time he entered the Stuttgart grammar school, where he remained for his education until he was 18. As a schoolboy he made a collection of extracts, alphabetically arranged, comprising annotations on classical authors, passages from newspapers, and treatises on morals and mathematics from the standard works of the period.

University and private studies

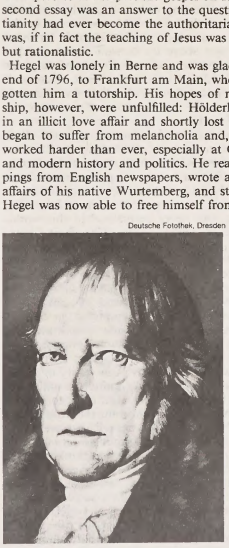
In 1788 Hegel went as a student to Tübingen with a view to taking orders, as his parents wished. Here he studied philosophy and classics for two years and graduated in 1790. Though he then took the theological course, he was impatient with the orthodoxy of his teachers; and the certificate given to him when he left in 1793 states that, whereas he had devoted himself vigorously to philosophy, his industry in theology was intermittent. He was also said to be poor in oral exposition, a deficiency that was to dog him throughout his life. Though his fellow students called him "the old man," he liked cheerful company and a "sacrifice to Bacchus" and enjoyed the ladies as well. His chief friends during that period were a pantheistic poet, J.C.F. Hölderlin, his contemporary, and the nature philosopher Schelling, five years his junior. Together they read the Greek tragedians and celebrated the glories of the French Revolution.

On leaving college, Hegel did not enter the ministry; instead, wishing to have leisure for the study of philosophy and Greek literature, he became a private tutor. For the next three years he lived in Berne, with time on his hands and the run of a good library, where he read Edward Gibbon on the fall of the Roman empire and *De l'esprit des loix*, by Charles Louis, baron de Montesquieu, as well as the Greek and Roman classics. He also studied the critical philosopher Immanuel Kant and was stimulated by his essay on religion to write certain papers that became noteworthy only when, more than a century later, they were published as a part of *Hegels theologische Jugendschriften* (1907). Kant had maintained that, whereas orthodoxy requires a faith in historical facts and in doctrines that reason alone cannot justify and imposes on the faithful a moral system of arbitrary commands alleged to be revealed, Jesus, on the contrary, had originally taught a rational morality, which was reconcilable with the teaching of Kant's ethical works, and a religion that, unlike Judaism, was adapted to the reason of all men. Hegel accepted this teaching; but, being more of a historian than Kant was, he put it to the test of history by writing two essays. The first of these was a life of Jesus in which Hegel attempted to reinterpret the gospel on Kantian lines. The second essay was an answer to the question of how Christianity had ever become the authoritarian religion that it was, if in fact the teaching of Jesus was not authoritarian but rationalistic.

Move to Frankfurt am Main

Hegel was lonely in Berne and was glad to move, at the end of 1796, to Frankfurt am Main, where Hölderlin had gotten him a tutorship. His hopes of more companionship, however, were unfulfilled: Hölderlin was engrossed in an illicit love affair and shortly lost his reason. Hegel began to suffer from melancholia and, to cure himself, worked harder than ever, especially at Greek philosophy and modern history and politics. He read and made clippings from English newspapers, wrote about the internal affairs of his native Wurtemberg, and studied economics. Hegel was now able to free himself from the domination

Deutsche Fotothek, Dresden



Hegel, oil painting by Jakob von Schlesinger, c. 1825. In the Staatliche Museen zu Berlin.

of Kant's influence and to look with a fresh eye on the problem of Christian origins.

Emancipation from Kantianism. It is impossible to exaggerate the importance that this problem had for Hegel. It is true that his early theological writings contain hard sayings about Christianity and the churches; but the object of his attack was orthodoxy, not theology itself. All that he wrote at this period throbs with a religious conviction of a kind that is totally absent from Kant and Hegel's other 18th-century teachers. Above all, he was inspired by a doctrine of the Holy Spirit. The spirit of man, his reason, is the candle of the Lord, he held, and therefore cannot be subject to the limitations that Kant had imposed upon it. This faith in reason, with its religious basis, henceforth animated the whole of Hegel's work.

His outlook had also become that of a historian—which again distinguishes him from Kant, who was much more influenced by the concepts of physical science. Every one of Hegel's major works was a history; and, indeed, it was among historians and classical scholars rather than among philosophers that his work mainly fructified in the 19th century.

When in 1798 Hegel turned back to look over the essays that he had written in Berne two or three years earlier, he saw with a historian's eye that, under Kant's influence, he had misrepresented the life and teachings of Jesus and the history of the Christian Church. His newly won insight then found expression in his essay "Der Geist des Christentums und sein Schicksal" ("The Spirit of Christianity and Its Fate"), likewise unpublished until 1907. This is one of Hegel's most remarkable works. Its style is often difficult and the connection of thought not always plain, but it is written with passion, insight, and conviction.

"The Spirit of Christianity"

He begins by sketching the essence of Judaism, which he paints in the darkest colours. The Jews were slaves to the Mosaic Law, leading a life unlovely in comparison with that of the ancient Greeks and content with the material satisfaction of a land flowing with milk and honey. Jesus taught something entirely different. Men are not to be the slaves of objective commands; the law is made for man. They are even to rise above the tension in moral experience between inclination and reason's law of duty, for the law is to be "fulfilled" in the love of God, wherein all tension ceases and the believer does God's will wholeheartedly and single-mindedly. A community of such believers is the Kingdom of God.

This is the kingdom that Jesus came to teach. It is founded on a belief in the unity of the divine and the human. The life that flows in them both is one; and it is only because man is spirit that he can grasp and comprehend the Spirit of God. Hegel works out this conception in an exegesis of passages in the Gospel According to John. The kingdom, however, can never be realized in this world: man is not spirit alone but flesh also. "Church and state, worship and life, piety and virtue, spiritual and worldly action can never dissolve into one."

In this essay the leading ideas of Hegel's system of philosophy are rooted. Kant had argued that man can have knowledge only of a finite world of appearances and that, whenever his reason attempts to go beyond this sphere and grapple with the infinite or with ultimate reality, it becomes entangled in insoluble contradictions. Hegel, however, found in love, conceived as a union of opposites, a prefigurement of spirit as the unity in which contradictions, such as infinite and finite, are embraced and synthesized. His choice of the word *Geist* to express this his leading conception was deliberate: the word means "spirit" as well as "mind" and thus has religious overtones. Contradictions in thinking at the scientific level of Kant's "understanding" are indeed inevitable, but thinking as an activity of spirit or "reason" can rise above them to a synthesis in which the contradictions are resolved. All of this, expressed in religious phraseology, is contained in the manuscripts written toward the end of Hegel's stay in Frankfurt. "In religion," he wrote, "finite life rises to infinite life." Kant's philosophy had to stop short of religion. But there is room for another philosophy, based on the concept of spirit, that will distill into conceptual form the insights of religion. This was the philosophy that Hegel now felt himself ready to expound.

Career as lecturer at Jena. Fortunately, his circumstances changed at this moment, and he was at last able to embark on the academic career that had long been his ambition. His father's death in 1799 had left him an inheritance, slender, indeed, but sufficient to enable him to surrender a regular income and take the risk of becoming a *Privatdozent*. In January of 1801 he arrived in Jena, where Schelling had been a professor since 1798. Jena, which had harboured the fantastic mysticism of the Schlegel brothers and their colleagues and the Kantianism and ethical Idealism of Fichte, had already seen its golden age, for these great scholars had all left. The precocious Schelling, who was but 26 on Hegel's arrival, already had several books to his credit. Apt to "philosophize in public," Schelling had been fighting a lone battle in the university against the rather dull followers of Kant. It was suggested that Hegel had been summoned as a new champion to aid his friend. This impression received some confirmation from the dissertation by which Hegel qualified as a university teacher, which betrays the influence of Schelling's philosophy of nature, as well as from Hegel's first publication, an essay entitled "Differenz des Fichte'schen und Schelling'schen Systems der Philosophie" (1801), in which he gave preference to the latter. Nevertheless, even in this essay and still more in its successors, Hegel's difference from Schelling was clearly marked; they had a common interest in the Greeks, they both wished to carry forward Kant's work, they were both iconoclasts; but Schelling had too many romantic enthusiasms for Hegel's liking; and all that Hegel took from him—and then only for a very short period—was a terminology.

Relationship with Schelling

Hegel's lectures, delivered in the winter of 1801–02, on logic and metaphysics, were attended by about 11 students. Later, in 1804, with a class of about 30, he lectured on his whole system, gradually working it out as he taught. Notice after notice of his lectures promised a textbook of philosophy—which, however, failed to appear. After the departure of Schelling from Jena (1803), Hegel was left to work out his own views untrammelled. Besides philosophical and political studies, he made extracts from books, attended lectures on physiology, and dabbled in other sciences. As a result of representations made by himself at Weimar, he was in February 1805 appointed extraordinary professor at Jena; and in July 1806, on Goethe's intervention, he drew his first stipend—100 thalers. Though some of his hearers became attached to him, Hegel was not yet a popular lecturer.

Hegel, like Goethe, felt no patriotic shudder when Napoleon won his victory at Jena (1806): in Prussia he saw only a corrupt and conceited bureaucracy. Writing to a friend on the day before the battle, he spoke with admiration of the "world soul" and the Emperor and with satisfaction at the probable overthrow of the Prussians.

At this time Hegel published his first great work, the *Phänomenologie des Geistes* (1807; Eng. trans., *The Phenomenology of Mind*, 2nd ed., 1931). This, perhaps the most brilliant and difficult of Hegel's books, describes how the human mind has risen from mere consciousness, through self-consciousness, reason, spirit, and religion, to absolute knowledge. Though man's native attitude toward existence is reliance on the senses, a little reflection is sufficient to show that the reality attributed to the external world is due as much to intellectual conceptions as to the senses and that these conceptions elude a man when he tries to fix them. If consciousness cannot detect a permanent object outside itself, so self-consciousness cannot find a permanent subject in itself. Through aloofness, scepticism, or imperfection, self-consciousness has isolated itself from the world; it has closed its gates against the stream of life. The perception of this is reason. Reason thus abandons its efforts to mold the world and is content to let the aims of individuals work out their results independently.

The Phenomenology of Mind

The stage of *Geist*, however, reveals the consciousness no longer as isolated, critical, and antagonistic but as the indwelling spirit of a community. This is the lowest stage of concrete consciousness, the age of unconscious morality. But, through increasing culture, the mind gradually

emancipates itself from conventions, which prepares the way for the rule of conscience. From the moral world the next step is religion. But the idea of Godhead, too, has to pass through nature worship and art before it reaches a full utterance in Christianity. Religion thus approaches the stage of absolute knowledge, of "the spirit knowing itself as spirit." Here, according to Hegel, is the field of philosophy.

GYMNASIUM RECTOR

In spite of the *Phänomenologie,* however, Hegel's fortunes were now at their lowest ebb. He was, therefore, glad to become editor of the *Bamberger Zeitung* (1807–08). This, however, was not a suitable vocation, and he gladly accepted the rectorship of the Aegidiengymnasium in Nürnberg, a post he held from December 1808 to August 1816 and one that offered him a small but assured income. There Hegel inspired confidence in his pupils and maintained discipline without pedantic interference in their associations and sports.

In 1811 Hegel married Marie von Tucher (22 years his junior), of Nürnberg. The marriage was entirely happy. His wife bore him two sons: Karl, who became eminent as a historian; and Immanuel, whose interests were theological. The family circle was joined by Ludwig, a natural son of Hegel's from Jena. At Nürnberg in 1812 appeared *Die objektive Logik,* being the first part of his *Wissenschaft der Logik* ("Science of Logic"), which in 1816 was completed by the second part, *Die subjecktive Logik.*

UNIVERSITY PROFESSOR

This work, in which his system was first presented in what was essentially its ultimate shape, earned him the offer of professorships at Erlangen, at Berlin, and at Heidelberg.

Hegel's philosophical system

At Heidelberg. He accepted the chair at Heidelberg. For use at his lectures there, he published his *Encyklopädie der philosophischen Wissenschaften im Grundrisse* (1817; "Encyclopaedia of the Philosophical Sciences in Outline"), an exposition of his system as a whole. Hegel's philosophy is an attempt to comprehend the entire universe as a systematic whole. The system is grounded in faith. In the Christian religion God has been revealed as truth and as spirit. As spirit, man can receive this revelation. In religion the truth is veiled in imagery; but in philosophy the veil is torn aside, so that man can know the infinite and see all things in God. Hegel's system is thus a spiritual monism but a monism in which differentiation is essential. Only through an experience of difference can the identity of thought and the object of thought be achieved—an identity in which thinking attains the through-and-through intelligibility that is its goal. Thus, truth is known only because error has been experienced and truth has triumphed; and God is infinite only because he has assumed the limitations of finitude and triumphed over them. Similarly, man's Fall was necessary if he was to attain moral goodness. Spirit, including the Infinite Spirit, knows itself as spirit only by contrast with nature. Hegel's system is monistic in having a single theme: what makes the universe intelligible is to see it as the eternal cyclical process whereby Absolute Spirit comes to knowledge of itself as spirit (1) through its own thinking; (2) through nature; and (3) through finite spirits and their self-expression in history and their self-discovery, in art, in religion, and in philosophy, as one with Absolute Spirit itself.

"Encyclopaedia of the Philosophical Sciences"

The compendium of Hegel's system, the "Encyclopaedia of the Philosophical Sciences," is in three parts: "Logic," "Nature," and "Mind." Hegel's method of exposition is dialectical. It often happens that in a discussion two people who at first present diametrically opposed points of view ultimately agree to reject their own partial views and to accept a new and broader view that does justice to the substance of each. Hegel believed that thinking always proceeds according to this pattern: it begins by laying down a positive thesis that is at once negated by its antithesis; then further thought produces the synthesis. But this in turn generates an antithesis, and the same process continues once more. The process, however, is circular: ultimately, thinking reaches a synthesis that is identical with its starting point, except that all that was implicit there has now been made explicit. Thus, thinking itself, as a process, has negativity as one of its constituent moments, and the finite is, as God's self-manifestation, part and parcel of the infinite itself. This is the sort of dialectical process of which Hegel's system provides an account in three phases.

"Logic." The system begins with an account of God's thinking "before the creation of nature and finite spirit"; *i.e.,* with the categories or pure forms of thought, which are the structure of all physical and intellectual life. Throughout, Hegel is dealing with pure essentialities, with spirit thinking its own essence; and these are linked together in a dialectical process that advances from abstract to concrete. If a man tries to think the notion of pure Being (the most abstract category of all), he finds that it is simply emptiness; *i.e.,* Nothing. Yet Nothing *is.* The notion of pure Being and the notion of Nothing are opposites; and yet each, as one tries to think it, passes over into the other. But the way out of the contradiction is at once to reject both notions separately and to affirm them both together; *i.e.,* to assert the notion of becoming, since what becomes both is and is not at once. The dialectical process advances through categories of increasing complexity and culminates with the absolute idea, or with the spirit as objective to itself.

"Nature." Nature is the opposite of spirit. The categories studied in "Logic" were all internally related to one another; they grew out of one another. Nature, on the other hand, is a sphere of external relations. Parts of space and moments of time exclude one another; and everything in nature is in space and time and is thus finite. But nature is created by spirit and bears the mark of its creator. Categories appear in it as its essential structure, and it is the task of the philosophy of nature to detect that structure and its dialectic; but nature, as the realm of externality, cannot be rational through and through, though the rationality prefigured in it becomes gradually explicit when man appears. In man nature rises to self-consciousness.

"Mind." Here Hegel follows the development of the human mind through the subconscious, consciousness, and the rational will; then through human institutions and human history as the embodiment or objectification of that will; and finally to art, religion, and philosophy, in which finally man knows himself as spirit, as one with God and possessed of absolute truth. Thus, it is now open to him to think his own essence; *i.e.,* the thoughts expounded in "Logic." He has finally returned to the starting point of the system, but en route he has made explicit all that was implicit in it and has discovered that "nothing but spirit is, and spirit is pure activity."

Hegel's system depends throughout on the results of scientific, historical, theological, and philosophical inquiry. No reader can fail to be impressed by the penetration and breadth of his mind nor by the immense range of knowledge that, in his view, had to precede the work of philosophizing. A civilization must be mature and, indeed, in its death throes before, in the philosophic thinking that has implicitly been its substance, it becomes conscious of itself and of its own significance. Thus, when philosophy comes on the scene, some form of the world has grown old.

The Philosophy of Right

At Berlin. In 1818 Hegel accepted the renewed offer of the chair of philosophy at Berlin, which had been vacant since Fichte's death. There his influence over his pupils was immense, and there he published his *Naturrecht und Staatswissenschaft im Grundrisse,* alternatively entitled *Grundlinien der Philosophie des Rechts* (1821; Eng. trans., *The Philosophy of Right,* 1942). In Hegel's works on politics and history, the human mind objectifies itself in its endeavour to find an object identical with itself. *The Philosophy of Right* (or *of Law*) falls into three main divisions. The first is concerned with law and rights as such: persons (*i.e.,* men as men, quite independently of their individual characters) are the subject of rights, and what is required of them is mere obedience, no matter what the motives of obedience may be. Right is thus an abstract universal and therefore does justice only to the universal element in the human will. The individual, however, cannot be satisfied unless the act that he does accords not merely with law but also with his own conscientious convictions. Thus, the problem in the modern world is to construct a social

and political order that satisfies the claims of both. And thus no political order can satisfy the demands of reason unless it is organized so as to avoid, on the one hand, a centralization that would make men slaves or ignore conscience and, on the other hand, an antinomianism that would allow freedom of conviction to any individual and so produce a licentiousness that would make social and political order impossible. The state that achieves this synthesis rests on the family and on the guild. It is unlike any state existing in Hegel's day; it is a form of limited monarchy, with parliamentary government, trial by jury, and toleration for Jews and dissenters.

After his publication of *The Philosophy of Right*, Hegel seems to have devoted himself almost entirely to his lectures. Between 1823 and 1827 his activity reached its maximum. His notes were subjected to perpetual revisions and additions. It is possible to form an idea of them from the shape in which they appear in his published writings. Those on *Aesthetics*, on the *Philosophy of Religion*, on the *Philosophy of History*, and on the *History of Philosophy* have been published by his editors, mainly from the notes of his students, whereas those on logic, psychology, and the philosophy of nature have been appended in the form of illustrative and explanatory notes to the corresponding sections of his *Encyklopädie*. During these years hundreds of hearers from all parts of Germany and beyond came under his influence; and his fame was carried abroad by eager or intelligent disciples.

Three courses of lectures are especially the product of his Berlin period: those on aesthetics, on the philosophy of religion, and on the philosophy of history. In the years preceding the revolution of 1830, public interest, excluded from political life, turned to theatres, concert rooms, and picture galleries. At these Hegel became a frequent and appreciative visitor, and he made extracts from the art notes in the newspapers. During his holiday excursions, his interest in the fine arts more than once took him out of his way to see some old painting. This familiarity with the facts of art, though neither deep nor historical, gave a freshness to his lectures on aesthetics, which, as put together from the notes taken in different years from 1820 to 1829, are among his most successful efforts.

Philosophies of aesthetics; of religion; of history

The lectures on the philosophy of religion are another application of his method, and shortly before his death he had prepared for the press a course of lectures on the proofs for the existence of God. On the one hand, he turned his weapons against the Rationalistic school, which reduced religion to the modicum compatible with an ordinary worldly mind. On the other hand, he criticized the school of Schleiermacher, who elevated feeling to a place in religion above systematic theology. In his middle way, Hegel attempted to show that the dogmatic creed is the rational development of what was implicit in religious feeling. To do so, of course, philosophy must be made the interpreter and the superior discipline.

In his philosophy of history, Hegel presupposed that the whole of human history is a process through which mankind has been making spiritual and moral progress and advancing to self-knowledge. History has a plot, and the philosopher's task is to discern it. Some historians have found its key in the operation of natural laws of various kinds. Hegel's attitude, however, rested on the faith that history is the enactment of God's purpose and that man had now advanced far enough to descry what that purpose is: it is the gradual realization of human freedom.

The first step was to make the transition from a natural life of savagery to a state of order and law. States had to be founded by force and violence; there is no other way to make men law-abiding before they have advanced far enough mentally to accept the rationality of an ordered life. There will be a stage at which some men have accepted the law and become free, while others remain slaves. In the modern world man has come to appreciate that all men, as minds, are free in essence, and his task is thus to frame institutions under which they will be free in fact.

Hegel did not believe, despite the charge of some critics, that history had ended in his lifetime. In particular, he maintained against Kant that to eliminate war is impossible. Each nation-state is an individual; and, as Hobbes had said of relations between individuals in the state of nature, pacts without the sword are but words. Clearly, Hegel's reverence for fact prevented him from accepting Kant's Idealism.

The lectures on the history of philosophy are especially remarkable for their treatment of Greek philosophy. Working without modern indexes and annotated editions, Hegel's grasp of Plato and Aristotle is astounding, and it is only just to recognize that it was from Hegel that the scholarship lavished on Greek philosophy in the century after his death received its original impetus.

Recognition and honours

At this time a Hegelian school began to gather. The flock included intelligent pupils, empty-headed imitators, and romantics who turned philosophy into lyric measures. Opposition and criticism only served to define more precisely the adherents of the new doctrine. Though he had soon resigned all direct official connection with the schools of Brandenburg, Hegel's real influence in Prussia was considerable. In 1830 he was rector of the university. In 1831 he received a decoration from Frederick William III. One of his last literary undertakings was the establishment of the Berlin *Jahrbücher für wissenschaftliche Kritik* ("Yearbook for Philosophical Criticism").

The revolution of 1830 was a great blow to Hegel, and the prospect of mob rule almost made him ill. His last literary work, the first part of which appeared in the *Preussische Staatszeitung* while the rest was censored, was an essay on the English Reform Bill of 1832, considering its probable effects on the character of the new members of Parliament and the measures that they might introduce. In the latter connection he enlarged on several points in which England had done less than many continental states for the abolition of monopolies and abuses.

In 1831 cholera entered Germany. Hegel and his family retired for the summer to the suburbs, and there he finished the revision of the first part of his *Science of Logic*. Home again for the winter session, on November 14, after one day's illness, he died of cholera and was buried, as he had wished, between Fichte and Karl Solger, author of an ironic dialectic.

PERSONAGE AND INFLUENCE

In his classroom Hegel was more impressive than fascinating. His students saw a plain, old-fashioned face, without life or lustre—a figure that had never looked young and was now prematurely aged. Sitting with his snuffbox before him and his head bent down, he looked ill at ease and kept turning the folios of his notes. His utterance was interrupted by frequent coughing; every sentence came out with a struggle. The style was no less irregular: sometimes in plain narrative the lecturer would be specially awkward, while in abstruse passages he seemed especially at home, rose into a natural eloquence, and carried away the hearer by the grandeur of his diction.

The early theological writings and the *Phenomenology of Mind* are packed with brilliant metaphors. In his later works, produced as textbooks for his lectures, the "Encyclopaedia of the Philosophical Sciences" and the *Philosophy of Right*, he compresses his material into relatively short, numbered paragraphs. It is only necessary to translate them to appreciate their conciseness and precision. The common idea that Hegel's is a philosophy of exceptional difficulty is quite mistaken. Once his terminology is understood and his main principles grasped, he presents far less difficulty than Kant, for example. One reason for this is a certain air of dogmatism: Kant's statements are often hedged around with qualifications; but Hegel had, as it were, seen a vision of absolute truth, and he expounds it with confidence.

Emergence of Hegelianism

Hegel's system is avowedly an attempt to unify opposites—spirit and nature, universal and particular, ideal and real—and to be a synthesis in which all the partial and contradictory philosophies of his predecessors are alike contained and transcended. It is thus both Idealism and Realism at once; hence, it is not surprising that his successors, emphasizing now one and now another strain in his thought, have interpreted him variously. Conservatives and revolutionaries, believers and atheists alike have professed to draw inspiration from him. In one form or

another his teaching dominated German universities for some years after his death and spread to France and to Italy. The vicissitudes of Hegelian thought to the present day are detailed below in *Hegelianism*. In the mid-20th century, interest in the early theological writings and in the *Phänomenologie* was increased by the spread of Existentialism. At the same time, the growing importance of Communism encouraged political thinkers to study Hegel's political works, as well as his "Logic," because of their influence on Karl Marx. And, by the time of his bicentennial in 1970, a Hegelian renascence was in the making. (T.M.K.)

Hegelianism

Hegelianism is the name given to a diversified philosophical movement that developed out of Hegel's monumental system of thought. The term is here so construed as to exclude Hegel himself and to include, therefore, only the ensuing Hegelian movements. As such, its thought is focussed upon history and logic, a history in which it sees, in various perspectives, that "the rational is the real" and a logic in which it sees that "the truth is the Whole."

GENERAL CONSIDERATIONS

Problems of the Hegelian heritage. The Hegelian system, in which German Idealism reached its fulfillment, claimed to provide a unitary solution to all of the problems of philosophy. It held that the speculative point of view, which transcends all particular and separate perspectives, must grasp the *one* truth, bringing back to its proper centre all of the problems of logic, of metaphysics (or the nature of Being), and of the philosophies of nature, law, history, and culture (artistic, religious, and philosophical). According to Hegel, this attitude is more than a formal method that remains extraneous to its own content; rather, it represents the actual development of the Absolute—of the all-embracing totality of reality—considered "as Subject and not merely as Substance" (*i.e.*, as a conscious agent or Spirit and not merely as a real being). This Absolute, Hegel held, first puts forth (or posits) itself in the immediacy of its own inner consciousness and then negates this positing—expressing itself now in the particularity and determinateness of the factual elements of life and culture—and finally regains itself, through the negation of the former negation that had constituted the finite world.

Dialectical development of the Absolute

Such a dialectical scheme (immediateness–alienation–negation of the negation) accomplished the self-resolution of the aforementioned problem areas—of logic, of metaphysics, and so on. This panoramic system thus had the merit of engaging philosophy in the consideration of all of the problems of history and culture, none of which could any longer be deemed foreign to its competence. At the same time, however, the system deprived all of the implicated elements and problems of their autonomy and particular authenticity, reducing them to symbolic manifestations of the one process, that of the Absolute Spirit's quest for and conquest of its own self. Moreover, such a speculative mediation between opposites, when directed to the more impending problems of the time, such as those of religion and politics, led ultimately to the evasion of the most urgent and imperious ideological demands and was hardly able to escape the charge of ambiguity and opportunism.

Stages in the history of the interpretation of Hegel. The explanation of the success of Hegelianism—marked by the formation of a school that, for more than 30 years, brought together the best energies of German philosophy—lies in the fact that no other system could compete with it in the richness of its content or the rigour of its formulation or challenge its claim to express the total spirit of the culture of its time. Moreover, as Hegelianism diffused outward, it was destined to provoke increasingly lively and gripping reactions and to take on various articulations as, in its historical development, it intermingled with contrasting positions.

Four stages can be distinguished within the development of Hegelianism. The first of these was that of the immediate crisis of the Hegelian school in Germany during the period from 1827 through 1850. Always involved in polemics against its adversaries, the school soon divided into three currents: (1) the right, in which the direct disciples of Hegel participated, defended his philosophy from the accusation that it was liberal and pantheistic (defining God as the All). These "old Hegelians" sought to uphold the compatibility of Hegelianism with evangelical orthodoxy and with the conservative political policies of the Restoration (the new order in Europe that followed the defeat of Napoleon). (2) The left—formed of the "young Hegelians," for the most part indirect disciples of Hegel—considered the dialectic as a "principle of movement" and viewed Hegel's identification of the rational with the real as a command to modify the cultural and political reality that reactionism was merely justifying and to make it rational. Thus the young Hegelians interpreted Hegelianism in a revolutionary sense—*i.e.*, as pantheistic and then, consecutively, as atheistic in religion and as liberal democratic in politics. (3) The centre, which preferred to fall back upon interpretations of the Hegelian system in its genesis and significance, with special interest in logical problems.

Hegelian right, centre, and left

In the second phase (1850–1904), in which Hegelianism diffused into other countries, the works of the centre played a preponderant role; thus in this phase of the history of the interpretation of Hegel, usually called Neo-Hegelian, the primary interest was in logic and a reform of the dialectic.

In the first decade of the 20th century, on the other hand, there arose still in Germany a different movement, after Wilhelm Dilthey, originator of a critical approach to history and humanistic studies, discovered unpublished papers from the period of Hegel's youth. This third phase, that of the Hegel renaissance, was characterized by an interest in philology, by the publication of texts, and by historical studies; and it stressed the reconstruction of the genesis of Hegel's thought, considering especially its cultural matrices—both Enlightenment and Romanticist—and the extent to which it might present irrationalistic and so-called pre-Existentialist attitudes.

In the fourth stage, after World War II, the revival of Marxist studies in Europe finally thrust into the foreground the interest in Hegel–Marx relationships and in the value of the Hegelian heritage for Marxism, with particular regard to political and social problems. This fourth phase of the history of Hegelianism thus appropriated many of the polemical themes of the earlier years of the school.

CRISES IN THE EARLIER HEGELIAN SCHOOL

The earlier development of Hegelianism can be divided, according to predominant concerns, into three periods: (1) polemics during the life of Hegel (1816–31), (2) controversies in the religious field (1831–39), and (3) political debates (1840–44), though discussions on all of the problems continued through all three periods.

Polemics during the life of Hegel: 1816–31. While Hegel was still living, discussion was dominated by the master. It was not a matter of polemics within the school but only one of objections against the system from various quarters: from speculative theists; from Johann Herbart, a prominent student of the philosophy of mind, and his followers; and from disciples of Friedrich Schelling, an objective and aesthetic Idealist, and of Friedrich Schleiermacher, a seminal thinker of modern theology.

The substantive history of the school stems from Hegel's later teaching at Berlin and from the publication of his *Naturrecht und Staatswissenschaft im Grundrisse* (1821; Eng. trans., *The Philosophy of Right*, 1942). This book was reviewed by Herbart, who reprimanded Hegel for mixing the monism of the Rationalist Spinoza with the transcendentalism of Kant, which had explored the conditions of the possibility of knowledge in general. There were also certain critics who directed the liberal press against Hegel for attacking Jakob Fries, a psychologizing Neo-Kantian, in the introduction of *The Philosophy of Right*. Some of the polemical writings of Hegel made a notable impact—*e.g.*, a preface that he wrote for a book by one of his earliest disciples, Hermann Hinrichs, on the relation of faith to reason (1822). In this preface, Hegel saw the two things

as the same in content but different in form—which for faith is the representation and for reason is the concept.

Articles in *Jahrbücher*

Particularly significant were eight articles in the *Jahrbücher für wissenschaftliche Kritik* (founded 1827; "Yearbooks for Scientific Critique"), a journal of the Hegelian right. Important among these were a review by Hegel that was unexpectedly eulogistic about the thesis that philosophy and evangelical orthodoxy are compatible and another review in which Hegel responded indirectly to arguments of Herbart. Among Hegel's critics can be distinguished speculative theists such as Christian Weisse of Leipzig and Immanuel Fichte, the son of the more famous Johann Fichte, who reproached him for his panlogism and proposed to unify thought and experience in the concept of a free God, the Creator. Among the most loyal disciples of Hegel were Hermann Hinrichs, his collaborator, and Karl Rosenkranz, who defended the Hegelian solution of the faith–reason problem (which had asserted the identity of content and difference of form), thus aptly defending the free rationality of religion.

Period of controversies chiefly in religion: 1831–39. The tone of these early polemics became animated and embittered after the death of Hegel. But, inasmuch as conditions in Germany, during the Restoration, inhibited the liberalization of political discussions, the milieu of controversy shifted to the religious realm and became related to problems of immortality, Christology, and general theology.

Feuerbach and Strauss

Shortly before Hegel's death, the youthful Ludwig Feuerbach, who later became a pioneer of naturalistic humanism, had published his *Gedanken über Tod und Unsterblichkeit* (1830; "Thoughts on Death and Immortality"), in which he contended that, from the Hegelian point of view, death must be necessary in order for man to be transformed from the finite to the infinite and it is thus a privilege for man preferable to empirical personal survival. This work was held to confirm the charge of pantheism that orthodox adversaries had directed at Hegel's system. On this point, at the appearance of two volumes by Johann Friedrich Richter, a pantheist and critic of religion, Hegel's disciples intervened, in an argument employing not a few dialectical artifices, to conciliate Hegelian statements with the traditional doctrine of immortality.

The polarization of historical positions that the debate on immortality could not adequately express soon came into the open with *Das Leben Jesu kritisch bearbeitet* (1835–36; Eng. trans., *The Life of Jesus Critically Examined*, 1846), of David Friedrich Strauss, a biblical interpreter and radical theologian. This work brought the problem of the nature of Christ up to date from the point of view that had been reached by biblical criticism; *i.e.*, Christology was no longer an issue of denominational dogma but, rather, a problem of the interpretation and evaluation of the Gospel sources and of their meaning in the historical development of civilization. In this approach, the narrowly philological outlook was overcome by a reconstruction in terms of a philosophy of history strangely suggestive of the young Hegel. The thesis of the book was that the Gospel account is interwoven with myths that are not the works of individuals but of the collective poetic activity of the first Christian community, myths that resulted in part from messianic expectations, in part from the memory of the historical figure of Jesus, and in part from a transfiguration of the real elements. The aim of the myths was to demonstrate that philosophy and religion are the same in content and to offer, in an imaginative guise (as in parables), the meaning of the one truth that Substance is unification of the divine nature and of the human, which Christ symbolized and which is realized in the spirit of all humanity.

Strauss's work provoked a lively reaction, to which he replied in his *Streitschriften* (1837–38; "Controversial Writings"), proposing the image of a Hegelian school split, like the French Parliament, into a right (Göschel, and several others), a centre (Rosenkranz), and a left (Strauss himself). There were responses from the right and centre and from Bruno Bauer, a philosopher, historian, and biblical critic. From the anti-Hegelian side there was, above all, *Die evangelische Geschichte* (1838; "The History of the Gospels"), by Weisse, who, conceding to Strauss the necessity to rationalize the Gospel story, propounded a speculative interpretation of the Christ figure as an incarnation of the Logos (Thought-Word), in contrast to the mystic and pantheistic views.

Weisse, Ruge, Schelling

Meanwhile, Bauer shifted toward the left in a polemic against the orthodox Ernst Hengstenberg, a vehement accuser of the Hegelians, and in his *Kritik der Geschichte der Offenbarung* (1838; "Critique of the History of Revelation"). In 1838 was founded the earliest journal of the left, the *Hallische Jahrbücher für deutsche Wissenschaft und Kunst* ("Halle Yearbooks for German Science and Art"), coedited by the activist philosopher Arnold Ruge and T. Echtermeyer. At first, the journal maintained a moderate tone, and Hegelians of the centre and right also contributed articles. In June, however, it veered to the democratic-liberal side as Ruge struck out against an accuser of the young Hegelians and as Feuerbach attacked earlier Hegelians. Hegelianism, which marks the culmination of speculative philosophy, Feurebach charged, does not demonstrate its own truth, because its contrast between sensory reality and intellectual concept comprises an irresoluble contradiction. Thus, its dialectic turns out to be a "monologue with itself," bereft of authentic mediation with the world. Hegelian philosophy, he held, is a "rational mystique," and what is needed is a return to nature, which, as objective reason, ought to become a principle of philosophy and of art. Thus an extensive examination of contemporary culture was conducted by the journal's editors in an article that depicted Romanticism as a movement degraded to a reactionary stance and extolled the spirit of reform and of liberal (yet loyalist) Prussianism.

As for issues in the fields of logic and metaphysics, after several polemical exchanges the interest of philosophers was attracted to the publicist reawakening that came to Schelling, who reactivated certain anti-Hegelian criticisms. These criticisms dealt with the impossibility of building a valid philosophy upon the pure concept assumed as a point of departure and endowed with autonomous movement. Such a philosophy would be vitiated by presuppositions of what ought to be demonstrated and by hypostatizations (*i.e.*, the making of an idea into an entity). Schelling proposed, on the other hand, that the real itself be taken as the subject of development, to be grasped with a "lively intuition"; and that, while accepting a "negative philosophy" (such as that of Rationalism and Hegel) pointing to the conditions *without which* one cannot think, one must also add a "positive philosophy" delineating the conditions *by means of which* thought and reality can exist, premised on the existence of a free creative God.

Period of atheistic and political radicalism: 1840–44. The ensuing years marked one of the most intense periods in the cultural life of modern Europe.

Anti-Hegelian criticism. Advancing from Aristotelian presuppositions, an important critique against the Hegelian logic was presented by the classical philosopher and philologist Friedrich Adolf Trendelenburg in his *Logische Untersuchungen* (1840; "Logical Investigations"). In Hegel's view, the passage from Being to Nothing and to Becoming can be posited as a pure beginning "without presuppositions" of logic. In Trendelenburg's view, however, this passage is vitiated by its spurious dependence upon the surreptitious presupposition of the Empirical movement, without which support neither the passage from Being to Nothing (and vice versa) nor the recognition of Becoming as the "truth" of this primal opposition of concepts can be justified. Secondly, he charged that Hegel confused (1) the logical opposition or contradiction of *A* against non-*A* with (2) the real contradiction or contrariety of *A* against *B*. Contradiction (1) consists in the mere repetition of the first term with a negative sign; and from it no concrete movement can proceed. In contrariety (2), however, the opposition of the second term to the first is concrete—thus the second term cannot be deduced from the first and, instead, should be derived on its own account from empirical experience. Thus Hegel constructed his entire system, Trendelenburg charged, on an arbitrary dialectic of elements intrinsically real (contraries), which he mistakenly treated as though they were abstract opposites (contradictories) and were such by logical necessity.

Meanwhile, Schelling continued to teach his "positive philosophy"—of mythology and of revelation (of a personal God). Hence the philosophy of the later Schelling became the target of all of the criticisms from the left and likewise exerted a notable influence on the speculative theists. Meanwhile, the centre, on account of the critique of Trendelenburg, oriented itself toward the future reforms of Hegelianism.

Kierkegaard and Fischer

Among those who attended Schelling's lectures was Søren Kierkegaard, the man who was destined to become one of the founding fathers of Existentialism and whose religious individualism represents the earliest major result of the diffusion of Hegelianism outside of Germany. In all of his works—but above all in his *Philosophiske Smuler* (1844; Eng. trans., *Philosophical Fragments*, 1936) and his *Afsluttende uvidenskabelig Efterskrift* (1846; Eng. trans., *Concluding Unscientific Postscript*, 1941)—Kierkegaard waged a continuous polemic against the philosophy of Hegel. He regarded Hegel as motivated by the spirit of the harmonious dialectical conciliation of every opposition and as committed to imposing universal and panlogistic resolutions upon the authentic antinomies of life. Kierkegaard saw these antinomies as emerging from the condition of the individual, as a single person, who, finding himself always stretching to attain ascendance over his existential limitations in his absorption in God and at the same time always thrust back upon himself by the incommensurability of this relationship, cannot find his salvation except through the paradoxical inversion of the rational values of speculative philosophy and through the "leap of faith" in the crucified Christ. Kierkegaard's claim that the nexus of problems characterizing man's condition as an existing being is irreducible to any other terms lay at the very roots of Existentialism. It was destined to condition the critical relationship of this current of thought to Hegelianism throughout its subsequent history. Moreover, Kierkegaard's thought, which Kierkegaard did not know—still more than that of Strauss—seemed reminiscent of those problem areas explored in the young Hegel's religious thought—issues that were destined to appear only later when Hegel research would gain precise knowledge of the writings of Hegel's youth.

At this time the attitude of the centre was oriented toward reforms of the Hegelian system in the field of logic and historiography, as reflected especially in the emergence of Kuno Fischer, one of the foremost historians of philosophy. In the fundamental triad of the dialectic, as Fischer saw it, Being and Nothing are not equally static and neutralizing. The real movement does not interpose itself into their relationship because Being is here to be understood as the Being of thought, which, to the degree that it is a thinking of Nothing, possesses that dynamic surplus that becomes manifest in the moment of Becoming. It was in making responses to this view that the forthcoming Neo-Hegelian movement in Europe found some of its motivations.

Theological radicalism. In 1840 political conditions in Germany changed with the succession of the young Frederick William IV, whose minister began to repress the liberal press and summoned to Berlin in an anti-Hegelian capacity both Schelling and the conservative jurist F.J. Stahl, a stubborn critic of Hegel. Far from weakening the movement, however, these actions radicalized its revolutionary manifestations. Strauss, in *Die christliche Glaubenslehre* (1840–41; "The Christian Doctrine of Faith"), reaffirmed the opposition of philosophical pantheism to religious theism as a means of reunifying the finite and the infinite; and Feuerbach established a philosophical anthropology in his major work *Das Wesen des Christentums* (1841; Eng. trans., *The Essence of Christianity*, new ed. 1957), in which man reappropriates his essence, which he had alienated from himself by hypostatizing it in the idea of God. The essence of man is reason, will, and love; and these three faculties comprise the consciousness of the human species as a knowledge of the infinity that man must regain. Man must thus reverse the theological propositions that express the spurious objectification of his universality in God; for this objectification had been effected through the individual consciousness in its effort to surmount its limitations. Thus Feuerbach interpreted the Christian mysteries as symbols of the alienation of human properties absolutized as divine attributes, and he criticized the contradictions of theology that are found in such concepts as God, the Trinity, the sacraments, and faith. Man's reappropriation of his essence from such religious alienation is consummated in the "new religion" of humanity, of which the supreme principle is that "man is God to man."

Feuerbach and Bauer

To this period belong also the major critiques of Bruno Bauer on the Johannine (1840) and Synoptic (1841–42) Gospels. Differentiating his position from the pantheistic and mysticizing Substance of Strauss, Bauer held that the Gospels were not the unconscious product of the original community but a product of the self-consciousness of the Spirit in a given stage of its development. There followed two works specifically concerning Hegel, in which, feigning an orthodoxy from which he charged Hegel with atheism and radicalism, Bauer maintained, in the form of a parody, the revolutionary interpretation of Hegel that became customary in the current of the Hegelian left.

Sociopolitical radicalism. In the years 1841–43, the repressive measures of the government reached ever more decisive extremes: Bauer was debarred from teaching; Feuerbach did not even attempt to teach; and Ruge was enjoined to publish the *Hallische* in Prussia instead of Leipzig. (Actually, he transferred it to Dresden and changed its name to the *Deutsche Jahrbücher*.) Here also appeared one of Ruge's major writings, "Die Hegelsche Rechtsphilosophie und die Politik unserer Zeit" (1842; "The Hegelian Philosophy of Right and the Politics of our Time"), in which Ruge denounced Hegel's political conservatism, charging that his contemplative reason was reduced to the acceptance of existing conditions, to the exclusion of every effort to modify reality, and to the absolutizing of the Prussian state as the model of an ideal state. Ruge's journal was suppressed early in 1843, but in March he published in Switzerland his *Anekdota zur neuesten deutschen Philosophie und Publicistik* ("Anecdotes for the Latest German Philosophy and Political Journalism"), containing articles by Bauer, Ruge, Marx, Feuerbach, and others.

Feuerbach's article developed the claim that the method of speculative philosophy, which is the ultimate form of theology, is to invert the subject and predicate—*i.e.*, to substantialize the abstract and to treat concrete determinations as attributes or "logical accidents" of hypostatized abstractions. The inversion of speculative propositions, he held, leads to the philosophical reappropriation of man's essence; the philosophy of the future will achieve mastery through the negation of the Hegelian philosophy—and this is exactly what he entitled his forthcoming book: *Grundsätze der Philosophie der Zukunft* (1843; "Basic Principles of the Philosophy of the Future"). In place of the immediate Absolute of Hegel, he argued, there must be substituted the immediate individual existent—corporeal, sensible, and rational. Man's reappropriation of himself will be possible whenever his need to transcend his own limitations finds fulfillment in another person and in the totality of the human species: "thus man is the measure of reason."

Meanwhile, a schism had been ripening in the left wing: (1) On the one hand, there were the "Free Berliners" (initially the young Friedrich Engels, later to become Marx's theoretician, the radical anarchist Max Stirner, and the Bauer brothers), who, deeming themselves faithful to Hegel, developed a philosophy of self-consciousness (understood in a subjective and superindividualistic sense) directed toward treating social and historical problems with aristocratic intellectual detachment. (2) On the other hand, there was the group that included Ruge, the publicist Moses Hess, the scholarly poet Heinrich Heine, and Karl Marx. Influenced in their theories by Feuerbach, this group directed radicalism toward an experience deepened by the classical Enlightenment and embraced the rising Socialism. They thus involved Hegel in their critique of the political, cultural, and philosophical conditions of the time. The most widely known result of the first trend was Stirner's book *Der Einzige und sein Eigentum* (1845; "The

Individual and His Property"), in which the fundamental thesis of individualistic anarchism can be discerned. The unique entity, in Stirner's view, is the individual, who must rebel against the attempt made by every authority and social organization to impose upon him a cause not his own and must be regarded as a focus of absolutely free initiative—a goal to be reached by emancipating himself from every idea-value imposed by tradition.

The work of Marx

The years between 1840 and 1844, however, saw the emergence of a figure incomparably more representative of the crisis of German Hegelianism than any already cited, that of Karl Marx, who was destined to guide the experience of this crisis toward a revolution of world historical scope. Marx's study of Hegel dates from his university years in Berlin, the earliest result of which was his doctoral dissertation with the exceedingly important preparatory notes, in which he ventured an original application of Hegelian method to the problem of the great crises in the history of philosophy. At first a friend of Bauer, Marx clung closely, however, to the democratic wing of the left. In 1843 he completed an important critical study of Hegel's *Philosophy of Right,* in which he reproached Hegel for having absolutized into an ideal state the Prussian state of the time. Such absolutizing, he charged, lent itself to generalizations of broad critical scope with respect to the idealistic procedure of hypostatizing the Idea and brought about (as allegorical derivatives from it) certain concrete political and social determinations, such as family, classes, and the state powers. Not yet a Communist, Marx nonetheless completed, in his *Kritik der hegelschen Staatsrechts* (written in the summer of 1843, published 1929; "Critique of Hegel's Constitutional Law"), a criticism of the erroneous relationship initiated in Hegel between society and the state, which was destined to lead Marx from the criticism of the modern state to that of modern society and its alienation.

It will be recalled that Hegel had likewise proposed the concept of alienation, describing the dialectic as a movement of the Absolute that was determined by its alienating and then regaining itself (thus overcoming the self-negation). Already in the *Economic and Philosophic Manuscripts of 1844* (German ed., 1932; Eng. trans., 1959), Marx had enunciated a general critique of the Hegelian dialectic that revealed its a priori nature, which, in Marx's view, was mystifying and alienated inasmuch as Hegel did nothing but sanction, by a method inverted with respect to real relationships, the alienation of all the concrete historical and human determinations.

Marx then directed himself against his former colleagues on the left—against Bauer in his *Die heilige Familie* (1845; Eng. trans., *The Holy Family,* 1956) and against Stirner in his *Die deutsche Ideologie* (1845–46; Eng. trans., *The German Ideology,* 1938), criticizing their "ideologism" (*i.e.,* the illusion that Idealism can be carried into the revolutionary camp since it is ideas that make history). The historical Materialism that Marx counterposed against Idealism expressed the conviction that the basis comprising the relations of production, both economic and social, conditions the superstructure of political, juridical, and cultural institutions and that the interchange among these spheres of production within the totality of an historical epoch must be designed to overcome their contradictions. This Materialism, though not belonging any more to Hegelianism, was destined nonetheless to remain linked to it by continuing polemical relationships and overlapping problem areas throughout the subsequent history of the movement.

Along with Marx must, of course, be mentioned his colleague Friedrich Engels, who was more tied, however, to the Hegelian conception of the dialectic—particularly regarding the dialectic of nature—than Marx was.

HEGELIANISM THROUGH THE 20TH CENTURY

Decline in Germany

Development and diffusion of Hegelianism in the later 19th century. In Germany, the second half of the 19th century witnessed a decline in the fortunes of Hegelianism, beginning with the *Hegel und seine Zeit* (1857; "Hegel and His Age"), by Rudolph Haym, a historian of the modern German spirit. The decline was urged on by Neo-Kantianism and Positivism as well as by the political realism of Bismarck. Hegelian influences still appeared in the first representatives of historicism (which urged that all things be viewed in the perspective of historical change). The surviving Hegelians, however, such as Kuno Fischer and Johann Erdmann, devoted themselves to the history of philosophy. Strauss and the Bauer brothers were won over to conservatism, and even Ruge, returning from exile in England, became a conservative.

Political and cultural problems: East Europe and the United States. The diffusion of Hegelianism outside of Germany was oriented in two directions. With respect to its political and cultural problems, the Hegelian experience developed in east European philosophers and critics such as the Polish count Augustus Cieszkowski, a religious thinker whose philosophy of action was initially influenced by the left; and the theistic metaphysician Bronislaw Trentowski. Among the Russians can be cited the literary critic Vissarion Belinsky, the democratic revolutionary writers Aleksandr Herzen and Nikolay Chernyshevsky, and certain anarchists such as the Russian exile and revolutionist Mikhail Bakunin. And among the French there were Hegelian Socialists such as Pierre-Joseph Proudhon.

St. Louis and Cincinnati schools

In the United States, the interest in Hegelianism was stimulated by its political aspects and its philosophy of history. Its two centres, the St. Louis and Cincinnati schools, seemed to duplicate the German schism between a conservative and a revolutionary tendency. The former was represented by the Hegelians of the St. Louis school: the German Henry Brokmeyer and the New Englander William Harris, a pedagogue and politician, and the circle that they founded called the St. Louis Philosophical Society, which published an influential organ, *The Journal of Speculative Philosophy.* Their legitimism, or support for legitimate sovereignty, was expressed in the quest for a foundation, dialectical as well as speculative, for American democracy and in a dialectical interpretation of the history of the United States. The Cincinnati group, on the other hand, gathered around August Willich, a former Prussian officer, and John Bernard Stallo, an organizer of the Republican Party. Willich had participated in the Revolution of 1848 as a democratic partisan in south Germany, and, as an exile, had been in lively intercourse with Marx. He founded the *Cincinnati Republikaner,* in which he reviewed Marx's *Zur Kritik der politischen Ökonomie* (1859) and endeavoured to base the principles of social democracy upon the humanistic foundations of Feuerbach. Stallo, on the other hand, tried to interpret the political philosophy of Hegel in republican terms. The democratic community became, for him, the realization of the dialectic rationality of the Spirit with a rigorous separation of church and state.

Logic and Metaphysics problems: Italy, England. The second trend in non-German Hegelianism was directed, in Italy and in England, to problems of logic and metaphysics. A vigorously speculative rethinking of the foundations of Hegel's *Wissenschaft der Logik* was engaged in by the major liberal Italian philosopher Bertrando Spaventa and his associates. Spaventa's *Studi sull' etica di Hegel* (1869) consisted of a direct liberal translation of Hegel's *Philosophy of Right.* Seeking to rediscover the connection between the thinking of the Italians of the 16th century and that of the German Idealists, Spaventa encountered the system of problems involved in the relationship between Kant and Hegel. He adopted from Kuno Fischer the solutions by which Fichte, Schelling, and Hegel had rendered Kant's transcendental ego consummatively veritable. He thus proposed an epistemological (idealistic theory of knowledge) interpretation of the Hegelian logic, according to which one premise of the logic is the dialectic of consciousness described in Hegel's *Phenomenology of Mind,* and the problems of the genesis of logic are resolved in the sense that Being is, from first to last, Becoming; *i.e.,* it is thought in action, which negates the objective residue of thought-out Being and, for that reason, is confirmed as a creative process. From Spaventa, whose intention was to vindicate the freedom and autonomy of thought against denominational dogmatism, was derived the foundation for the subjectivistic formalization

of Hegelianism soon undertaken by Giovanni Gentile, an early-20th-century Idealist.

As in Italy, so also in England, interest in Hegel arose from the philoopher's need to round out his experience of classical German thought by tracing its vicissitudes since the time of Kant; and this interest was directed toward the fields of epistemology and logic and in this instance was applied to problems of religion and not of politics. The pioneer in English Hegelianism was James Hutchison Stirling, through his work *The Secret of Hegel* (1865). Stirling reaffirmed the lineage of thought that Fischer had traced "from Kant to Hegel," endeavouring to penetrate the dialectic-speculative relationship of unity in multiplicity as the central point of the dialectic. Toward Hegelianism as a unifying experience the ethics scholar Thomas Hill Green, the foremost representative of Hegelianism at Oxford, applied himself, though with more original attitudes; and the brothers John and Edward Caird dedicated themselves to right-wing interpretations of religious subjects—Edward in a well-known monograph entitled *Hegel* (1883).

Hegelianism in the first half of the 20th century. At this point, the development of Hegelianism branched out in two directions: one of which, in England and Italy, pursued the tendencies of the Neo-Hegelians of the preceding decades, while the other, in Germany and France, accomplished the philological interpretative renewal known as the Hegel renaissance.

Neo-Hegelianism in England and Italy. With respect to the first tendency, there appeared in England at the turn of the century various outstanding works on Hegel's logic by authors who were partly Hegelian in spirit. These scholars, toiling through the system of problems that they shared—which focussed on establishing a criterion for the unification of the multiplicity of experience—ended up in diverse positions: those of Bernard Bosanquet and John Ellis MacTaggart, for example, who were translators and commentators of Hegelian works; but above all that of the foremost spiritualistic philosopher then in England, F.H. Bradley, author of the renowned *Appearance and Reality* (1893), whose development led him to positions more and more at odds with the absolute panlogism of Hegel. His affirmation of the dualism of appearance and reality was the result of a critique of the category of relations, which, by introducing contradictions between the qualities of the thing, utterly shattered the unity of experience in which it might seem that true reality could be reached—a reality that in Bradley's view it is not given to thought to attain.

Bradley, Royce, Croce

The echoes of this Idealistic system were not long in being felt in the United States by one of its most profound philosophers, an absolute Idealist, Josiah Royce, who, in *The World and the Individual* (1900–01), discussed the skeptical Idealism of Bradley in order to overthrow its consequences in favour of a conception of the infinite as a self-representative system and of the world (or the All) as an individualized realization of the intentional aims of the Idea copresent in a superior eternal consciousness. In Anglo-Saxon Neo-Hegelianism, the Hegelian experience has always been merely an episode—which fact serves to refine, by contrast, the methods of experimentalism that are more congenial to the Empirical tradition in England.

In Italy, on the other hand, the Neo-Hegelianism of the 20th century took the form of a spiritualistic reaction to the spread of Positivism that had followed upon the unification of Italy. This reaction developed in two directions: that of the historicism of Benedetto Croce and that of the actualism of Giovanni Gentile, two scholars who divided the realm of philosophy between themselves and occupied it—rather heavy-handedly—for four decades. The Crocean reform of Hegelianism dates from his volume *Ciò che è vivo e ciò che è morto della filosofia di Hegel* (1907; "What Is Living and What Is Dead in the Philosophy of Hegel") and from the systematic works of his so-called "philosophy of the spirit." Croce accepted the dialectic from Hegel as a requirement for the unification of opposites; but he rejected its system, in which Hegel would put in opposition and treat dialectically certain intellectual forms that are not really opposite but only distinct—such as the beautiful, the true, the useful, and the good, each of which has its dialectical opposite over against itself that it has to overcome within the purview of each grade. Consequently, renouncing the possibility of a philosophy of nature or of history, Croce formulated a development of so-called "distinct grades" according to the spiritual forms of art, of philosophy, of economics, and of ethics and contended that the comprehensive meaning of the development of the Spirit is given by history "as thought and as action" and a realization of freedom.

Gentile, on the other hand, accentuated the opposition of subject and object by considering every objective factuality as surpassed by the living dialectical development of the act—*i.e.*, the becoming of the Spirit in its own self-making, proceeding from an originating self-establishment, or *autoktisis*, of the Spirit itself. From this position he derived an absolute subjectivism that exploited all the possibilities for dialectically transforming every fixed position into its opposite, a downright sophistry of disengagement. Gentile's pro-Fascist stance, however, condemned his actualism to collapse.

Hegelian renaissance in Germany and France. Already from the beginnings of the century, however, there had been in Germany a change in Hegelian interpretation instigated by Wilhelm Dilthey's re-examination, in 1905, of the youthful manuscripts of Hegel and by the publication by one of Dilthey's principal disciples, Herman Nohl, of *Hegels theologische Jugendschriften* (1907; "The Theological Writings of Hegel's Youth"). Inasmuch as there had been heretofore only fragmentary notices on these unpublished literary remains, the effect of this rereading of the texts was to place them in contrast with the works of his maturity; they thus emerged as dealing, for the most part, with various problem areas in ethics, religion, and history; as lacking systematic preoccupations; and as rich discourse, tending to the mystic, which invited their comparison with the severe technical uniformity of his major works. Hermeneutical interest, however, centred especially on the problem of the beginnings of the philosophy and dialectic of Hegel, of which the first formulations were investigated in order to collate their meanings with those of the major works and of the *Phenomenology*, which was a key work of the Hegelian evolution inasmuch as it participated both in the romanticized colouring of the youthful writings and in the systematic demands of the *Encyklopädie der philosophischen Wissenschaften im Grundrisse* (1817; "Encyclopaedia of the Philosophical Sciences in Outline").

Publication of the *Jugendschriften*

Scholars were soon led to investigate the historical matrices of Hegel's intellectual culture—the late Enlightenment and dawning Romanticism—a direction of inquiry that yielded imposing contributions rich in discussions that continue to this day. These studies began with Dilthey's monograph, which pointed out the irrationalistic and vitalistic aspects of Hegel's youthful writings. In addition, a basic work by Franz Rosenzweig, *Hegel und der Staat* (1920), genetically reconstructed the political thought of the young Hegel in relation to its historical sources and concluded that the influence of Rousseau prevented Hegel from becoming the genuine "national philosopher of Germany." Jean Wahl, a French metaphysician and historian of philosophy, wrote on the "wretched conscience," interpreting Hegel existentially. Further, the German philosopher Richard Kroner studied the development from Kant to Hegel integrating it with the contributions of early Romanticism. And Hermann Glockner, a Bavarian aesthetic intuitionist, saw following one another in the development of Hegel a so-called "pantragistic" phase up to the *Phenomenology* and, subsequently, an opposing "panlogistic" phase that betrayed the most lively and concrete instances of the preceding phase—a work that approached the efforts at interpreting Hegel that were made by the Nazis.

Hegelian studies today. Today one has to speak not of the presence of Hegelianism as an operating philosophical current but only of studies on Hegel and of an experience of the Hegelian philosophy, to which, however, almost none of the present-day orientations in philosophy is foreign. The repeated encounter of Western culture with Marxist thought after World War II has brought to the fore the political, ethical, and religious implications of Hegelianism; and a marshalling into opposing camps analogous to that of the earlier crisis of the school is

taking shape. Today there are no orthodox Hegelians, but there are denominational critics of Hegelianism, especially Catholic, whose cognizance of Hegel's painful development invokes, despite their differences, a certain fellow feeling with him.

In the centre are found scholars of a liberal and radical frame of mind but with varying orientations with respect to historical interpretations. Karl Löwith, a German philosopher of history and culture, sees Hegel as the initiator of the "historicist" crisis in modern thought, culminating in Marx and in Kierkegaard; and to this he contrasts the metahistorical perspective reflected in the Nietzschean motif of the "eternal return," based on the ideal of a Goethean serenity. In France, Alexandre Kojève, noteworthy for his effort to harmonize Hegel with Martin Heidegger, proposes a reinterpretation of the *Phänomenologie* as a manifesto of the emancipation of "man the servant" from all alienations. Jean Hyppolite, author of an outstanding commentary on the *Phänomenologie,* usually presents a restrained humanistic interpretation of the Hegel of Jena. This renaissance of the study of Hegel has conditioned the thought of some of the major thinkers of France. Particularly notable, however, is the Hegelian conditioning of German philosopher-sociologists such as Theodor Adorno and Herbert Marcuse. The former is sometimes regarded as the most Hegelian thinker of the mid-20th century because he sought to bring again to the fore Hegel's dialectic, understood in a new anti-intellectualistic sense, as a method for the solution of present-day social problems. Marcuse, a partisan of a Diltheian interpretation, approaches the position of the first Hegelian left, ending up in what critics see as a neoromantic anarchism. The major merit of both of these thinkers lies in their incisive analyses of aspects of modern consumer societies, especially American—though their proposed remedies remain uncertain.

Three currents of Marxist interpretation

The major interest, however, in the contemporary interpretation of Hegel is displayed by the Marxist camp. Marxist interpretation of Hegel had permeated the entire history of Hegelianism (notwithstanding the fact that the critical activity of young Marx against Hegel had been vehemently conducted and had led to various effects). This interpretation had settled upon the distinction made by Friedrich Engels between the method and the system of Hegel's philosophy—*i.e.,* between the dialectic considered as a revolutionary "principle of movement" that achieves fulfillment in human culture, and the system, regarded, on the other hand, as reactionary because idealistic and conservative. With varying emphases on critical issues, this interpretation was continued in subsequent Marxist thinkers—from the Russians Georgy Plekhanov and Lenin to Mao Tse-tung and Joseph Stalin—the latter of whom affirmed the complementariness of historical and dialectical Materialism.

Today many Marxist scholars, especially in the countries of eastern Europe, remain favourable to the traditional line of Engels; and above all György Lukács, a Hungarian philosopher and literary critic and author of a volume on the young Hegel, does so. With the intention of revealing the romantic and irrationalistic presuppositions of Naziism, Lukács reevaluates, in German culture, the tendency of the Enlightenment and of democracy, which he recognizes in the young Goethe, in Schiller, in Hölderlin, and in the young Hegel—in whom he sees, however, a reactionary involution.

A secondary tendency, which is drawing attention in France, with the work of Louis Althusser, draws Marx close to Structuralism, a recent school that seeks through a "human science," to probe the systematic structures evinced in cultural life. In this school Marx's humanism is viewed as a temporary, Feuerbachian phase, surpassed by commitment to the scientific observation of the structure of bourgeois society. Such Structuralistic interpretation of Marxism thus runs the risk of departing from a due emphasis on the *historical* substance of Marxian Materialism.

The latter motive is, on the other hand, the essential aim of a third Marxist current, in Italy, initiated by Galvano della Volpe, a critical aesthetician who discusses the relationship between bourgeois and Socialist democracy and champions, in aesthetics, a critical and antiromantic Aristotelianism. This current has been continued by Mario Rossi, who asks one to read again in full the texts of Hegel and Marx, to reconstruct the related movements, and to compare the Materialistic conception of history with more recent philosophical currents such as Structuralism, present-day sociology, and the logic of the sciences.

A conclusion of a theoretical-systematic nature concerning Hegelianism has today become not only impossible but also inopportune, because its possible interest has been effectively replaced by that of the sheer history of the movement. The latter has shown how the substantial ambiguity of the philosophy and dialectic of Hegel can be resolved only when its claim to be able to solve all problems on a theoretical level and to achieve a "circular" decisiveness in its arguments—which violates the conditioning specificity of historical facts—is refuted. It is then the scholar's task to explore the limits of Hegel's thought as well as its conditioned inadequacies—but also its merits, which are above all those of having expressed and documented the major part of the cultural problems of modern civilization. (M.R.)

BIBLIOGRAPHY

Hegel. *Works:* A collected edition of Hegel's published works, together with a great deal of material culled from his lectures was published by his pupils within a few years of his death in 1831. This edition, with some rearrangement, was reissued by HERMANN GLOCKNER in 26 volumes, including a comprehensive index (1927–40). In 1905 the Philosophische Bibliothek (Leipzig, later Hamburg) began publication of a new edition with a carefully revised text edited by GEORG LASSON and later by JOHANNES HOFFMEISTER; volumes appeared for more than 50 years, but it was not completed. It has been enhanced by a comprehensive edition sponsored by the Deutsche Forschungsgemeinschaft, which is to contain about 50 volumes. The first volume appeared in 1968. English translations of most of Hegel's works were published in the late 19th and early 20th century, but, apart from those by WILLIAM WALLACE (*Logic* and *Mind—i.e.,* the first and third parts of the *Encyclopaedia*), they are not always satisfactory and they have no notes. With a view to remedying this deficiency, new English translations have appeared of some works, including *Philosophy of Right* (1942, often reprinted) and *Science of Logic* (1969), as well as translations of writings not translated previously, such as *Early Theological Writings* (1948; rev. ed., 1971), and *Philosophy of Nature,* 3 vol. (1970), the second part of the *Encyclopaedia.* With the exception of *Science of Logic* and the Oxford translation of *Philosophy of Nature,* all these translations are annotated.

Life and philosophy: RAYMOND PLANT, *Hegel* (1973), is a study of origins of his thought; and CHARLES TAYLOR, *Hegel* (1975), is a study of the development of his philosophy. An excellent short account of Hegel's philosophy in English is EDWARD CAIRD, *Hegel* (1883, reissued 1972), but it has been updated in certain respects by G.R.G. MURE, *The Philosophy of Hegel* (1965); and in more detail by WALTER A. KAUFMANN, *Hegel* (1965, reissued 1978). An attempt to interest modern philosophers in Hegel is contained in J.N. FINDLAY, *Hegel* (1958, reissued 1976), but this important and lively work is for consideration only by those already acquainted with Hegel. As an introduction, G.R.G. MURE, *An Introduction to Hegel* (1940, reprinted 1982), is more reliable but it is not an exposition. A standard long exposition of Hegel's mature system is KUNO FISCHER, *Hegels Leben, Werke und Lehre,* especially the 2nd ed. (1911, reprinted 1976); while in English there is WALTER T. STACE, *The Philosophy of Hegel* (1924). In 1900 Wilhelm Dilthey maintained that Hegel could be understood only if there were a study of his early manuscripts; on the basis of these, Dilthey wrote *Die Jugendgeschichte Hegels* (1906, reissued 1968), a history of Hegel's development. This seminal work, hardly noticed at all by writers in English before 1965, gave rise to an immense literature in Germany, France, and Italy. An important and brilliant study of the young Hegel is GEORG LUKACS, *Der junge Hegel* (1948; Eng. trans. 1975), written from a Marxist point of view. An exhaustive study is that of THEODOR HAERING, *Hegel: sein Wollen und sein Werk,* 2 vol. (1929–38, reissued 1979). Reliable and more readable than this are the two volumes of HERMANN GLOCKNER issued (1929 and 1940) as an appendix to his edition of the collected works. Even these, however, have been outdated by the flood of material collected by the Hegel-Archiv at Bochum in West Germany, and published in a series of volumes of *Hegel-Studien* (1961 and subsequent years), and by the four volumes of Hegel's letters, edited by JOHANNES HOFFMEISTER (1952–60). An admirable, but now obsolete, biography is by KARL ROSENKRANZ (1824).

Specialized commentaries: (*Mind*): JUDITH N. SHKLAR, *Freedom and Independence: A Study of the Political Ideas of Hegel's "Phenomenology of Mind"* (1976), is a guidebook and commentary. (*Phenomenology of Spirit*): JEAN HYPPOLITE, *Genèse et structure de la Phénoménologie de l'Esprit de Hegel* (1946, reissued 1970). (*Logic*): G.R.G. MURE, *A Study of Hegel's Logic* (1950, reissued 1967); STANLEY ROSEN, *G.W.F. Hegel: An Introduction to the Science of Wisdom* (1974), is a study of the development and meaning of his dialectic; and HANS-GEORG GADAMER, *Hegel's Dialectic: Five Hermeneutical Studies* (1976), is an analysis for specialists. (*Nature*): The apparatus in the English translation by M.J. PETRY, 3 vol. (1970), provides a full and learned commentary. (*Law, morality, and the state*): WILHELM SEEBERGER, *Hegel; oder, die Entwicklung des Geistes zur Freiheit* (1961), is a good introduction to Hegel's thought as a whole; but HUGH A. REYBURN, *The Ethical Theory of Hegel* (1921, reissued 1970); and FRANZ ROSENZWEIG, *Hegel und der Staat*, 2 vol. (1920, reissued 1962), are excellent summaries, and the latter is a commentary as well. SHLOMO AVINERI, *Hegel's Theory of the Modern State* (1972), is an account of his political thought. See also GEORGE A. KELLY, *Hegel's Retreat from Eleusis: Studies in Political Thought* (1978). (*Art*): JACK KAMINSKY, *Hegel on Art* (1962, reissued 1970), is a fair summary of Hegel's lectures. (*Religion*): THOMAS M. KNOX, *A Layman's Quest* (1969), deals, in chapters 5 and 6, not only with Hegel's lectures on the philosophy of religion but also with all his other writings elsewhere on religion; BERNARD M.G. REARDON, *Hegel's Philosophy of Religion* (1977), is a summary. (*History*): BURLEIGH T. WILKINS, *Hegel's Philosophy of History* (1974), is an introduction; and GEORGE D. O'BRIEN, *Hegel on Reason and History: A Contemporary Reinterpretation* (1975), questions his reputation as an anti-empirical, apriorist thinker.

Hegelianism. *Critical works:* Works presenting a critical consideration of Hegelianism viewed as a whole are few. See, however: STEPHAN D. CRITES, "Hegelianism," in the *Encyclopedia of Philosophy*, vol. 3, pp. 451–459 (1967, reissued 1972); MARIO ROSSI, *Da Hegel a Marx*, 2 vol. (1970); and RENÉ SERREAU, *Hegel et l'hégélianisme*, 4th ed. (1971).

Historical works: JOHN E. TOEWS, *Hegelianism: The Path Toward Dialectical Humanism, 1805–1841* (1980); JOHANN E. ERDMANN, *Die Deutsche Philosophie seit Hegels Tode* (1963); WILLY MOOG, *Hegel und die Hegelsche Schule* (1930, reissued 1973); KARL LÖWITH, *From Hegel to Nietzsche: The Revolution in Nineteenth-Century Thought*, 1964; originally published in German, 1941); and two anthologies, on the Left and Right, respectively: KARL LÖWITH (ed.), *Die Hegelsche Linke* (1962); and HERMANN LÜBBE (ed.), *Die Hegelsche Rechte* (1962).

In various countries: (*Germany*): HEINRICH LEVY, *Die Hegel-Renaissance in der deutschen Philosophie* (1927). (*Italy*): MARIO ROSSI (ed.), *Sviluppi dello Hegelismo in Italia* (1957); BENEDETTO CROCE, *Saggio sullo Hegel*, 5th ed. (1967). (*Slavic countries*): Contributions of authors from Russia, Poland, the Balkans, and Czechoslovakia are presented in *Hegel bei den Slaven*, 2nd ed., ed. by DMITRIJ TSCHIŽEWSKIJ (1961); see also BORIS JAKOWENKO, *Ein Beitrag zur Geschichte des Hegelianismus in Russland* (1934). (*England*): HIRA-LAL HALDAR, *Neo-Hegelianism* (1927). (*United States*): LOYD D. EASTON, "Hegelianism in Nineteenth-Century Ohio," *Journal of the History of Ideas*, 23:355–378 (1962), for the Cincinnati school; HENRY A. POCHMANN, *German Culture in America: Philosophical and Literary Influences 1600–1900*, pp. 257–294 (1957, reprinted 1978), for the St. Louis school. See also WILLIAM H. GOETZMANN and DICKSON PRATT (eds.), *The American Hegelians: An Intellectual Episode in the History of Western America* (1973).

Other works: AUGUSTE CORNU, *Karl Marx et Friedrich Engels*, 2 vol. (1955–58), is rich in materials and citations from the *Hallische* and *Deutsche Jahrbücher*. See also HERBERT MARCUSE, *Reason and Revolution: Hegel and the Rise of Social Theory*, 2nd ed. (1954); and SIDNEY HOOK, *From Hegel to Marx* (1936, reissued 1950 and 1962).

(T.M.K./M.R./Ed.)

Heisenberg

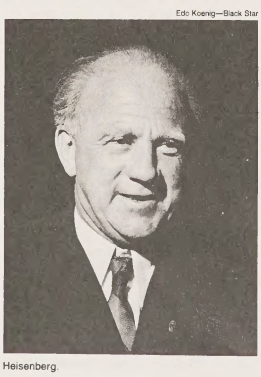
Werner Karl Heisenberg, physicist, philosopher, and public figure, helped to establish the modern science of quantum mechanics, out of which came the famous indeterminacy principle. He also made important contributions to the theories of the hydrodynamics of turbulence, the atomic nucleus, ferromagnetism, cosmic rays, and elementary particles; and he planned the first post-World War II German nuclear reactor, at Karlsruhe, West Germany.

Major influences

In his philosophical and methodological writings, Heisenberg was much influenced by Niels Bohr and Albert Einstein. From the former he derived the concepts of the social and dialogical character of scientific invention; the principle of correspondence (pragmatic and model-theoretical continuity) between macrophysics and microphysics; the permanence, though not the universality, of classical physics; the "interactive," rather than passive, role of the scientific observer in microphysics; and, consequently, the contextualized character of microphysical theories. From Einstein he derived the concepts of simplicity as a criterion of the central order of nature; scientific realism (*i.e.*, science describing nature itself, not merely how nature can be manipulated); and the theory-ladenness of scientific observations. He was coauthor with Bohr of the philosophy of complementarity. In his later work he conceived of a central order in nature, consisting of a set of universal symmetries expressible in a single mathematical equation for all systems of particulate matter. As a public figure, he actively promoted the peaceful use of nuclear energy after World War II and, in 1957, led other German scientists in opposing a move to equip the West German Army with nuclear weapons. He was, in 1954, one of the organizers of the Conseil Européen pour la Recherche Nucléaire (CERN; later, Organisation Européene pour la Recherche Nucléaire) in Geneva.

Early life. Heisenberg was born on December 5, 1901, in Würzburg, Germany. He studied physics, together with Wolfgang Pauli, his lifelong friend and collaborator, under Arnold Sommerfeld at the University of Munich and completed his doctoral dissertation (1923) on turbulence in fluid streams. Heisenberg followed Pauli to the University of Göttingen and studied there under Max Born; then, in the fall of 1924, he went to the Institute for Theoretical Physics in Copenhagen to study under Bohr.

Edo Koenig—Black Star



Heisenberg

Heisenberg's interest in Bohr's model of the planetary atom and his comprehension of its limitations led him to seek a theoretical basis for a new model. Bohr's concept—after 1913 the centrepiece of what has come to be called the old quantum theory—had been based on the classical motion of electrons in well-defined orbits around the nucleus, and the quantum restrictions had been imposed arbitrarily to bring the consequences of the model into conformity with experimental results. As a summary of existing knowledge and as a stimulus to further research, the Bohr atom had succeeded admirably, but the results of new research were becoming more and more difficult to reconcile with it.

In June 1925, while recuperating from an attack of hay fever on Helgoland, an island in the North Sea, Heisenberg solved a major physical problem—how to account for the stationary (discrete) energy states of an anharmonic oscillator. His solution, because it was analogous to that of a simple planetary atom, launched the program for the development of the quantum mechanics of atomic systems. (Quantum mechanics is the science that accounts for discrete energy states—as in the light of atomic spectra—and other forms of quantized energy, and for the phenomenon of stability exhibited by atomic systems.) Heisenberg published his results some months later in the *Zeitschrift für Physik* under the title "Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen" ("About the Quantum-Theoretical Reinterpretation of Kinetic and Mechanical Relationships"). In this article he proposed a reinterpretation of the basic concepts of mechanics.

Heisenberg's treatment of the problem departed from Bohr's as much as Bohr's had from 19th-century tenets. Heisenberg was willing to sacrifice the idea of discrete particles moving in prescribed paths (neither particles nor paths could be observed) in exchange for a theory that would deal directly with experimental facts and lead to the quantum conditions as consequences of the theory rather than ad hoc stipulations. Physical variables were to be represented by arrays of numbers; under the influence of Einstein's paper on relativity (1905), he took the variables to represent not hidden, inaccessible structures but "observable" (*i.e.*, measurable) quantities. Born saw that the arrays obeyed the rules of matrix algebra; he, Pascual Jordan, and Heisenberg were able to express the new theory in terms of this branch of mathematics, and the new quantum theory became matrix mechanics. Each (usually infinite-dimensional) matrix of the theory specified the set of possible values for a physical variable, and the individual terms of a matrix were taken to generate probabilities of occurrences of states and transitions among states. Heisenberg used the new matrix mechanics to interpret the dual spectrum of the helium atom (that is, the superposed spectra of its two forms, in which the spins of the two electrons are either parallel or antiparallel), and with it he predicted that the hydrogen molecule should have analogous dual forms. With others, he also addressed many atomic and molecular spectra, ferromagnetic phenomena, and electromagnetic behaviour. Important alternative forms of the new quantum theory were proposed in 1926 by Erwin Schrödinger (wave mechanics) and P.A.M. Dirac (transformation theory).

Indeterminacy principle

In 1927 Heisenberg published the indeterminacy, or uncertainty, principle. The form he derived appeared in a paper that tried to show how matrix mechanics could be interpreted in terms of the intuitively familiar concepts of classical physics. If q is the position coordinate of an electron (in some specified state), and p its momentum, assuming that q, and independently, p have been measured for many electrons (all in the particular state), then, Heisenberg proved,

$$\Delta q \cdot \Delta p > h,$$

where Δq is the standard deviation of measurements of q, Δp is the standard deviation of measurements of p, and h is Planck's constant (6.626176×10^{-27} erg-second). Indeterminacy principles are characteristic of quantum physics; they state the theoretical limitations imposed upon any pair of noncommuting (*i.e.*, conjugate) variables, such as the matrix representations of position and momentum; in such cases, the measurement of one affects the measurement of the other. The enormous significance of the indeterminacy principle is recognized by all scientists; but how it is to be understood physically—whether it depends on using intuitive classical ("complementary") pictures of a quantum system, or whether it is a principle in (a new kind of quantum) statistics, or whether in some sense through the special properties of the mathematical model it also describes a character of individual quantum systems—has been and still is much disputed. Bohr took the principle to apply to the complementary pictures of a quantum system—as a particle or as a wave pocket in classically intuited space; Heisenberg originally took the principle to apply to the nonintuitive properties of quantum, as distinct from classical, systems.

Bohr and Heisenberg elaborated a philosophy of complementarity to take into account the new physical variables and an appropriate measurement process on which each depends. This new conception of the measurement process in physics emphasized the active role of the scientist, who, in making measurements, interacted with the observed object and thus caused it to be revealed not as it is in itself but as a function of measurement. Many physicists, including Albert Einstein, Erwin Schrödinger, and Louis de Broglie, refused to accept the philosophy of complementarity.

Later life. From 1927 to 1941 Heisenberg was professor at the University of Leipzig. For the following four years, he was director of the Kaiser Wilhelm Institute for Physics in Berlin. Although he did not publicly oppose the Nazi regime, he was hostile to its policies. During World War II he worked with Otto Hahn, one of the discoverers of nuclear fission, on the development of a nuclear reactor. He failed to develop an effective program for nuclear weapons, probably from want of technical resources and lack of will to do so. After the war he organized and became director of the Max Planck Institute for Physics and Astrophysics at Göttingen, moving with the institute, in 1958, to Munich; he was also, in 1954, the German representative for the organizing of CERN.

Universal nonlinear spinor equation

In the postwar period Heisenberg began working on a fundamental spinor equation (a nonlinear differential equation capable of representing with spinors—complex vector-like entities—all possible particulate states of matter). His intuitions had led him to postulate that such an equation would exhibit a basic set of universal symmetries in nature (a symmetry is a mathematic form invariant under groups of canonical space–time and other changes in the representing elements), and be capable of explaining the variety of elementary particles generated in high-energy collisions. In this work, the "Platonic" character of which he recognized, he had the support and collaboration of Hans-Peter Dürr and Carl Friedrich von Weizsäcker.

Although he early, and indirectly, came under the influence of Ernst Mach, Heisenberg, in his philosophical writings about quantum mechanics, vigorously opposed the Logical Positivism developed by philosophers of science of the Vienna Circle. According to Heisenberg, what was revealed by active observation was not an absolute datum, but a theory-laden datum; *i.e.*, relativized by theory and contextualized by observational situations. He took classical mechanics and electromagnetics, which articulated the objective motions of bodies in space–time, to be permanently valid, though not applicable to quantum mechanical systems; he took causality to apply in general not to individual quantum mechanical systems but to mathematical representations alone, since particle behaviour could be predicted only on the basis of probability.

Heisenberg married Elisabeth Schumacher in 1937; they had seven children. He loved music in addition to physics and saw a deep affinity between these two interests. He also wrote philosophical works, believing that new insights into the ancient problems of Part and Whole and One and Many would help discovery in microphysics. Widely acknowledged as one of the seminal thinkers of the 20th century, Heisenberg was awarded the Nobel Prize for Physics for 1932 and was honoured with the Max Planck Medal, the Matteucci Medal, and the Barnard College Medal of Columbia University. He died in Munich on February 1, 1976. (P.A.H.)

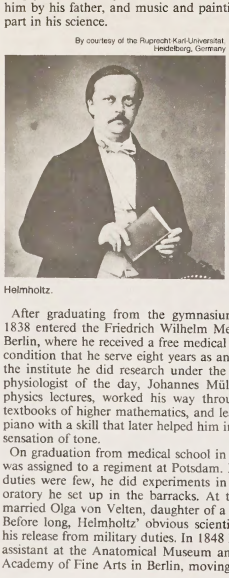
BIBLIOGRAPHY. Books by Heisenberg include *Physical Principles of the Quantum Theory* (1930, reprinted 1950; originally published in German, 1930), his most important work, containing themes of early papers amplified into a treatise; *Philosophic Problems of Nuclear Science* (1952, reprinted 1966; originally published in German, 1942, enlarged ed. 1948), a collection of his early essays; *Physics and Philosophy: The Revolution in Modern Science* (1958, reprinted 1962), his Gifford lectures; *Physics and Beyond* (1971, reprinted 1972; originally published in German, 1969), a memoir of his early life; and *Across the Frontiers* (1974; originally published in German, 1971), collected essays and occasional lectures. Biographical material is found in ARMIN HERMANN, *Werner Heisenberg, 1901–1976* (1976); CARL F. VON WEIZSÄCKER, *Werner Heisenberg* (1977); ELISABETH HEISENBERG, *Das politische Leben eines Unpolitischen: Erinnerungen an Werner Heisenberg* (1980), and the works cited there. For Heisenberg's role in the German wartime atomic program, see LESLIE R. GROVES, *Now It Can Be Told: Story of the Manhattan Project* (1962, reprinted 1983). Collections in honour of Heisenberg include FRITZ BOPP (ed.), *Werner Heisenberg und die Physik unserer Zeit* (1961); HEINRICH PFEIFFER (ed.), *Denken und Umdenken: Zu Werk und Wirkung von Werner Heisenberg* (1977); and PETER BREITENLOHNER and H. PETER DÜRR (eds.), *Unified Theories of Elementary Particles: Proceedings of the Heisenberg Symposium, July 16–21, 1981* (1982). For the history of quantum mechanics, see THOMAS KUHN *et al.* (eds.), *Sources for History of Quantum Physics* (1967), interviews with the principal originators of the quantum theory, the transcripts and tapes of which are deposited in many places; and EDWARD M. MACKINNON, *Scientific Explanation and Atomic Physics* (1982), a detailed historical account, with an excellent bibliography. For studies of Heisenberg's philosophy of science, see PATRICK A. HEELAN, *Quantum Mechanics and Objectivity* (1965); MAX JAMMER, *The Philosophy of Quantum Mechanics: The Interpretation of Quantum Mechanics in Historical Perspective* (1974), and *The Conceptual Development of Quantum Mechanics* (1966), which provide the most complete study of Heisenberg's contribution to quantum mechanics.

Helmholtz

One of the greatest scientists of the 19th century, Hermann Ludwig Ferdinand von Helmholtz made fundamental contributions to physiology, optics, electrodynamics, mathematics, and meteorology, but he is best known for his statement of the law of the conservation of energy. In addition, he brought to his laboratory research the ability to analyze the philosophical assumptions on which much of 19th-century science was based, and he did so with clarity and precision.

Early life. Helmholtz, the eldest of four children, was born at Potsdam, near Berlin, on August 31, 1821, and because of his delicate health was confined to home for his first seven years. His father was a teacher of philosophy and literature at the Potsdam Gymnasium, and his mother was descended from William Penn, the founder of Pennsylvania. From his mother came the calm and reserve that marked him all his life. From his father came a rich, but mixed, intellectual heritage. His father taught him the classical languages, as well as French, English, and Italian. He also introduced him to the philosophy of Immanuel Kant and Johann Gottlieb Fichte and to the approach to nature that flowed from their philosophical insights. This "Nature philosophy," in the hands of early 19th-century investigators, became a speculative science in which it was felt that scientific conclusions could be deduced from philosophical ideas, rather than from empirical data gathered from observations of the natural world. Much of Helmholtz' later work was devoted to refuting this point of view. His empiricism, however, was always deeply influenced by the aesthetic sensitivity passed on to him by his father, and music and painting played a large part in his science.

By courtesy of the Ruprecht-Karl-Universität, Heidelberg, Germany



Helmholtz.

After graduating from the gymnasium, Helmholtz in 1838 entered the Friedrich Wilhelm Medical Institute in Berlin, where he received a free medical education on the condition that he serve eight years as an army doctor. At the institute he did research under the greatest German physiologist of the day, Johannes Müller. He attended physics lectures, worked his way through the standard textbooks of higher mathematics, and learned to play the piano with a skill that later helped him in his work on the sensation of tone.

On graduation from medical school in 1843, Helmholtz was assigned to a regiment at Potsdam. Because his army duties were few, he did experiments in a makeshift laboratory he set up in the barracks. At that time he also married Olga von Velten, daughter of a military surgeon. Before long, Helmholtz' obvious scientific talents led to his release from military duties. In 1848 he was appointed assistant at the Anatomical Museum and lecturer at the Academy of Fine Arts in Berlin, moving the next year to Königsberg, in East Prussia (now Kaliningrad), to become assistant professor and director of the Physiological Institute. But Königsberg's harsh climate was injurious to his wife's health, and in 1855 he became professor of anatomy and physiology at the University of Bonn, moving in 1858 to Heidelberg. During these years his scientific interests progressed from physiology to physics. His growing scientific stature was further recognized in 1871 by the offer of the professorship of physics at the University of Berlin; in 1882, by his elevation to the nobility; and, in 1888, by his appointment as first director of the Physico-Technical Institute at Berlin, the post that he held for the rest of his life.

Early success

The variety of positions he held reflects his interests and competence but does not reflect the way in which his mind worked. He did not start out in medicine, move to physiology, then drift into mathematics and physics. Rather, he was able to coordinate the insights he had acquired from his experience in these disciplines and to apply them to every problem he examined. His greatest work, *Handbook of Physiological Optics* (1867), was characterized—like all of his scientific works—by a keen philosophical insight, molded by exact physiological investigations, and illustrated with mathematical precision and sound physical principles.

Attack on Nature philosophy

The general theme that runs through most, if not all, of Helmholtz' work may be traced to his rejection of Nature philosophy, and the violence of his rejection of this seductive view of the world may well indicate the early attraction it had for him. Nature philosophy derived from Kant, who in the 1780s had suggested that the concepts of time, space, and causation were not products of sense experience but mental attributes by which it was possible to perceive the world. Therefore, the mind did not merely record order in nature, as the Empiricists insisted; rather, the mind organized the world of perceptions so that, reflecting the divine reason, it could deduce the system of the world from a few basic principles. Helmholtz opposed this view by insisting that all knowledge came through the senses. Furthermore, all science could and should be reduced to the laws of classical mechanics, which, in his view, encompassed matter, force, and, later, energy, as the whole of reality.

Helmholtz' approach to nature was evident in the very first scientific researches he undertook while working for his doctorate in the laboratory of Müller. Like most biologists, Müller was a vitalist who was convinced that it would be impossible ever to reduce living processes to the ordinary mechanical laws of physics and chemistry. The organism as a whole, he insisted, was greater than the sum of its physiological parts. There must be some vital force that coordinated the physiological action of organs to produce the harmonious organic behaviour that characterized the living creature. Such a vital force was not susceptible to experimental investigation, and Müller therefore concluded that a truly experimental physiology was impossible.

Opposition to Müller's vitalism

In Müller's laboratory Helmholtz met a group of young men, among whom were Emil Heinrich Du Bois-Reymond, the founder of experimental neurophysiology, and Ernst Wilhelm von Brücke, who later became an expert on the operations of the human eye. Du Bois-Reymond expressed their opposition to Müller's views in a statement that fully expressed Helmholtz' own position. "Brücke and I," Du Bois-Reymond wrote, "we have sworn to each other to validate the basic truth that in an organism no other forces have any effect than the common physiochemical ones. . . ."

It was with this attitude that Helmholtz began his doctoral thesis in 1842 on the connection between nerve fibres and nerve cells. This soon led him to a broader

field of inquiry, namely, the source of animal heat. Recent publications in France had cast doubt upon the earlier confident assertion that all the heat produced in an animal body was the result of the heats of combination of the various chemical elements involved, particularly carbon, hydrogen, and oxygen. In 1842 Justus von Liebig attempted to reestablish the mechanical theory of animal heat in his book *Animal Chemistry; or, Organic Chemistry in Its Application to Physiology and Pathology.* Liebig tried to do this by experiments, whereas Helmholtz took a much more general path. Having mastered both physics and mathematics, Helmholtz could do what no other physiologist of the time could even attempt—subject the problem to a mathematical and physical analysis. He supposed that, if vital heat were not the sum of all the heats of the substances involved in chemical reactions within the organic body, there must be some other source of heat not subject to physical laws. This, of course, was precisely what the vitalists argued. But such a source, Helmholtz went on, would permit the creation of a perpetual motion machine if the heat could, somehow, be harnessed. Physics, however, had rejected the possibility of a perpetual motion machine as early as 1775, when the Paris Academy of Sciences had declared itself on the question. Hence, Helmholtz concluded, vital heat must be the product of mechanical forces within the organism. From there he went on to generalize his results to state that all heat was related to ordinary forces and, finally, to state that force itself could never be destroyed. His paper "On the Conservation of Force," which appeared in 1847, marked an epoch in both the history of physiology and the history of physics. For physiology, it provided a fundamental statement about organic nature that permitted physiologists henceforth to perform the same kind of material and energy balances as their colleagues in physics and chemistry. For the physical sciences, it provided one of the first, and certainly the clearest, statements of the principle of the conservation of energy.

In 1850 Helmholtz drove another nail into the coffin of vitalism. Müller had used the nerve impulse as an example of a vital function that would never be submitted to experimental measurement. Helmholtz found that this impulse was perfectly measurable and had the remarkably slow speed of some 90 feet (27 metres) per second. (This measurement was obtained by the invention of the myograph and illustrates Helmholtz' ability to create new instruments.) The slowness of the nerve impulse further supported those who insisted that it must involve the rearrangement of ponderable molecules, not the mysterious passage of a vital force.

Invention of the ophthalmoscope and ophthalmometer

Among Helmholtz' most valuable inventions were the ophthalmoscope and the ophthalmometer (1851). While doing work on the eye, and incidentally showing that it was a rather imperfect piece of workmanship not at all consonant with the vitalistic idea of the divine mind at work, Helmholtz discovered that he could focus the light reflected from the retina to produce a sharp image of the tissue. The ophthalmoscope remains one of the most important instruments of the physician, who can use it to examine retinal blood vessels, from which clues to high blood pressure and to arterial disease may be observed. The ophthalmometer permits the measurement of the accommodation of the eye to changing optical circumstances, allowing, among other things, the proper prescription of eyeglasses.

Helmholtz' researches on the eye were incorporated in his *Handbook of Physiological Optics,* the first volume of which appeared in 1856. In the second volume (1867), Helmholtz further investigated optical appearances and, more importantly, came to grips with a philosophical problem that was to occupy him for some years—Kant's insistence that such basic concepts as time and space were not learned by experience but were provided by the mind to make sense of what the mind perceived. The problem had been greatly complicated by Müller's statement of what he called the law of specific nerve energies. Müller discovered that sensory organs always "report" their own sense no matter how they are stimulated. Thus, for example, a blow to the eye, which has nothing whatsoever to do with optical phenomena, causes the recipient to "see stars." Obviously, the eye is not reporting accurately on the external world, for the reality is the blow, not the stars. How, then, is it possible to have confidence in what the senses report about the external world? Helmholtz examined this question exhaustively in both his work on optics and in his masterly *On the Sensation of Tone As a Physiological Basis for the Theory of Music* (1863). What he tried to do, without complete success, was to trace sensations through the sensory nerves and anatomical structures (such as the inner ear) to the brain in the hope of laying bare the complete mechanism of sensation. This task, it might be noted, has not been completed, and physiologists are still engaged in solving the mystery of how the mind knows anything about the outside world.

Helmholtz' detailed investigation of vision permitted him to refute Kant's theory of space by showing exactly how the sense of vision created the idea of space. Space, according to Helmholtz, was a learned, not an inherent, concept. Moreover, Helmholtz also attacked Kant's insistence that space was necessarily three-dimensional because that was how the mind had to conceive it. Using his considerable mathematical talents, he investigated the properties of non-Euclidean space and showed that these could be conceived and worked with as easily as the geometry of three dimensions.

Helmholtz' mathematical talents were not restricted to such theoretical planes as non-Euclidean geometry. He attacked and solved equations that had long frustrated physicists and mathematicians. In 1858 he published the paper "On the Integrals of Hydrodynamic Equations to Which Vortex Motions Conform." This was not only a mathematical tour de force, but, for a brief time, it also seemed to provide a key to the fundamental structure of matter. One of the consequences that flowed from Helmholtz' mathematical analysis was that vortices of an ideal fluid were amazingly stable; they could collide elastically with one another, intertwine to form complex knotlike structures, and undergo tensions and compressions, all without losing their identities. In 1866 William Thomson (later Lord Kelvin) proposed that these vortices, if composed of the ether that was presumed to be the basis for optical, electrical, and magnetic phenomena, could act exactly like primeval atoms of solid matter. Thus the ether would become the only substance in the cosmos, and all physical phenomena could be accounted for in terms of its static and dynamic properties.

Later life. Helmholtz' work in electricity and magnetism revealed his conviction that classical mechanics was probably the best mode of scientific reasoning. He was one of the first German scientists to appreciate the work in electrodynamics of the British scientists Michael Faraday and James Clerk Maxwell. Faraday had appeared to strike at the foundation of Newtonian physics by his unorthodox rejection of action at a distance, that is, action between two bodies in space without alteration of the medium between them. Maxwell, however, by interpreting the mathematics of Faraday's laws, showed there was no contradiction between Newtonian physics and classical mechanics. Helmholtz further developed the mathematics of electrodynamics. He spent his last years unsuccessfully trying to reduce all of electrodynamics to a minimum set of mathematical principles, an attempt in which he had to rely increasingly on the mechanical properties of the ether thought to pervade all space.

The mathematics of electrodynamics

Helmholtz was not in complete accord with Maxwell on the nature of electricity. Unlike Maxwell, Helmholtz was interested in and had studied electrochemistry, particularly the nature of the galvanic cell. Maxwell would have made the electric current solely the result of the polarization of the ether, or of whatever medium the current flowed through. Helmholtz, on the other hand, was fully conversant with Faraday's laws of electrolysis, which related the amount of current that passed through an electrochemical cell to the equivalent weights of the elements deposited at the poles. In 1881, in a lecture delivered in Faraday's honour in London, Helmholtz argued that if scientists accepted the existence of chemical atoms, as most chemists of the time did, then Faraday's laws necessarily implied

the particulate nature of electricity. This hypothetical particle was soon christened the electron and, ironically, the physics of its existence helped to falsify Helmholtz' theories of electrodynamics.

Though he was unsuccessful in his goal to formulate electrodynamics, Helmholtz was almost able to deduce all electromagnetic effects from the ether's supposed properties. The discovery of radio waves by his pupil Heinrich Hertz in 1888 was viewed as the experimental confirmation of the theories of Faraday, Maxwell, and Helmholtz. The special and general theories of relativity, proposed by Albert Einstein, destroyed Helmholtz' theories by eliminating the ether.

Helmholtz' early work on sound and music had led him to the study of wave motion. His work on the conservation of energy familiarized him with the problems of energy transfer. These two areas coalesced in his later years in his studies of meteorology, but the phenomena were so complex that he could do little more than point the way to future areas of research.

Helmholtz' work was the end product of the development of classical mechanics. He pushed it as far as it could go. When Helmholtz died in Berlin on September 8, 1894, the world of physics was poised on the brink of revolution. The discovery of X-rays, radioactivity, and relativity led to a new kind of physics in which Helmholtz' achievements, although impressive, had little to offer the new generation.

BIBLIOGRAPHY. There are two biographies of Helmholtz available to the reader of English. LEO KOENIGSBERGER, *Hermann von Helmholtz*, 3 vol. (1902–03; abridged Eng. trans., 1906, reprinted 1965), is often technical and sometimes difficult to understand. JOHN G. M'KENDRICK, *Hermann Ludwig Ferdinand von Helmholtz* (1899), deals only with Helmholtz' medical career. RICHARD M. and ROSLYN P. WARREN have published a collection of Helmholtz' writings on perception, entitled *Helmholtz on Perception: Its Physiology and Development* (1968), with critical comments. Helmholtz' *Popular Lectures on Scientific Subjects*, 2 vol. (1873; 2nd series, 1881), are excellent introductions to his thought.

(L.P.W.)

Heraldry

Heraldry is the science and art that deals with the use, display, and regulation of hereditary symbols employed to distinguish individuals, institutions, and corporations. These symbols, which probably originated as identification devices on shields, are called armorial bearings. Strictly defined, heraldry denotes that which pertains to the office and duty of a herald; that part of his work dealing with armorial bearings is properly termed armory. But in general usage, the term heraldry has come to mean the same as armory.

Like all other human creations, heraldic art has reflected the changes of fashion. As heraldry advanced from its utilitarian usages, its artistic quality declined. In the 18th century, for example, heraldry described new arms in an absurdly obtuse manner and rendered them in an overly intricate style. It was not until the 20th century that heraldic art recovered and in many ways improved upon the originals. There were still, however, far too many drawings of poor quality emanating from official sources.

This article is divided into the following sections:

The scope of heraldry 503
- General considerations 503
- The chief components of armorial bearings 505
- The elements and grammar of heraldic design 507
  - The field or ground of the shield
  - The "charges" on the field
  - The nature and origins of heraldic terminology
- The reading of heraldry 509
- Manipulation of heraldic design 510
  - Cadency
  - Arms of women
  - Quarterings and marshalling
  - Arms of bastardy
  - Nonfamilial heraldry

Historical development of heraldry 513
- Early roots of heraldry 513
- Growth of heraldry after the 13th century 515
  - Writers on heraldry
  - Continental versus British heraldry
- 20th-century heraldry 517
- Uses of heraldry for study and verification 518

Bibliography 518

The scope of heraldry

GENERAL CONSIDERATIONS

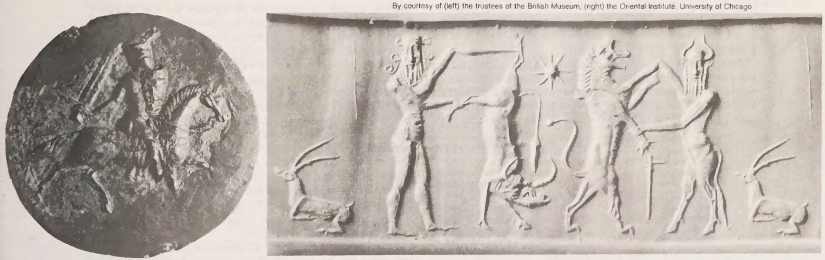
From the second quarter of the 12th century in western Europe, heraldic designs are found in general application. Elsewhere, a similar system is to be found only in Japan, in the *mon*, also dating from the 12th century. Other times and places are often said to have produced heraldic systems; for example, ancient Israel in the symbols of the 12 tribes, or the designs used by the Rajput princes in India. These and similar instances, however, are more properly considered incipient heraldry, since they did not develop into the complex heraldic practice known in western Europe and Japan.

From 1150 to 1500, the use of heraldry in the West was utilitarian: on armour in warfare, and on seals in peace. In the latter part of that period, it was used in peaceful ways and had much artistic value. Also, because from the beginning the use of arms had been associated with the higher feudal castes, heraldry acquired in later medieval times an identification with the concept of gentility that has persisted. To bear arms was the mark of a gentleman; therefore, to possess the desirable quality of gentility, a man needed to have armorial bearings. The great majority of those who seek to use coats of arms in the late 20th century are actuated by this motive. In the use of corporate arms, the motive of prestige rather than social distinction operates. As long as the possession of arms confers any social distinction, arms will be sought and used. At no previous time has there been so widespread an employment of heraldic devices.

The use of symbols has been universal among civilized communities, but these symbols have not assumed the character always associated with heraldry. Seals, too, which have a prominent place in heraldic practice, are of an antiquity approaching that of the most ancient civilizations. They were in use in the states that from Sumer onward flourished in Mesopotamia. Their use, for example, in the Babylonian Empire was the same as in medieval western Europe: to authenticate the documents (possibly of baked brick, later papyrus, later still parchment or vellum) on which they appeared or to which they were appended. All persons, literate and illiterate alike, were able to recognize the representation or symbol of a ruler or other potentate. In 12th-century Europe, heraldry first appeared on seals in the representations of persons. There is a clear line of

Seals as ancient devices of heraldry

By courtesy of (left) the trustees of the British Museum; (right) the Oriental Institute, University of Chicago



Seals as heraldic prototypes.
(Left) Impression of the great seal of Richard I of England, showing the mounted king bearing arms. In the British Museum. (Right) Cylinder seal impression from the Akkadian Period with a combat scene between a bearded hero and a bull-man, and beasts. In the Oriental Institute, University of Chicago.

descent from the seals of Assyria and Babylonia to the modern company seal, which is often heraldic.

Although originating in the small half continent of western Europe, heraldry has become universal, often, but not only, by way of western European colonization. Heraldry has spread to a considerable degree in both the Americas, Australia, New Zealand, and South Africa. In the former British India, the hereditary princes adopted the use of heraldry. In the numerous independent states formed in Africa from the former British colonies, official armorial bearings are generally used, and the same is true of the new states that were formerly French colonies. In Russia in the 18th century, the use of armorial bearings was adopted from the West, and state emblems are not unknown in Communist eastern Europe. In the 13th century, the Celtic princes of Wales and Ireland and the chiefs of the Scottish Highland clans took up the use of heraldic symbols from the example of the feudal lords and knights of other parts of Europe.

Other kinds of emblematic identification have some similarities with heraldry. An example is the totem system, found among the indigenous peoples of America and Australia, in which an animal, plant, or other object serves as an emblem of family or clan and is often regarded as a reminder of its ancestry. Totemism varies greatly in different countries, as do the theories that have been advanced to explain it. The totem poles used by the Indians of the northwest coast of North America contain a heraldic element in their employment of a hereditary symbol for a family or tribe. They therefore come under the heading of approaches to heraldic designs and may be termed semi-heraldic in character.

The Japanese *mon* as a type of armorial bearing

The Japanese *mon* is very definitely a heraldic symbol, having many parallels in its use with the armorial bearings of Europe. It was used on helmets, shields, and breastplates but never, as in Europe, large enough to identify the wearer of the armour at any considerable distance. When identification was desired, the *mon* was displayed on flags. The *mon* has usually been equated in English with "crest" and in some European languages has been translated erroneously as "coat of arms." It most closely resembles the heraldic badge (distinctive mark used by retainers), however, which in Europe often antedated armorial bearings. Further resemblances to European heraldry in the use of the *mon* include: the decorative use of the symbol on clothes, furniture, and houses; the use on the clothes of retainers of great lords; the legal requirement of registration of the *mon* (dating from the 17th century); and the reservation of the chrysanthemum *mon* to the emperor, with junior members of the imperial family using a different variety of the flower. This last distinction corresponds exactly to the rules of heraldic precedence that apply to the European royal families. That areas so far removed from each other as western Europe and Japan should have developed a system of hereditary symbolism independently of one another is not surprising, for in both areas feudalism was the prevailing medieval political and social system. As in Europe, Japanese heraldry survived the obsolescence of armour and has remained in widespread use in the 20th century.

Despite some uninformed opinion to the contrary, tartan has no connection with heraldry. It is simply a form of weaving cloth that is by no means restricted to the Scottish Highlands. Armorial bearings were adopted by Highland chiefs in imitation of the Lowland chivalry from the 13th and 14th centuries. The badge of the chief was adopted and used extensively by the members of his clan.

Drawing by Wm. A. Norman



The 16-petalled chrysanthemum *mon* of the Japanese Imperial family.

Flags can be heraldic. That of the United Kingdom is certainly so; it is formed by the amalgamation of the flags of England, Scotland, and Ireland, these showing respectively the crosses of St. George, St. Andrew, and St. Patrick, all of which are displayed heraldically. The United States flag has a quasi-heraldic character and appears to owe its principal ingredients to the armorial bearings of the first president, George Washington. The flag representing the republic of France, by contrast, is not heraldic, being merely an arrangement of the national colours.

In addition to national flags, there are banners, rectangular pieces of cloth showing the armorial bearings of the owner, and standards, strips of cloth that taper gradually to the end and usually bear heraldic devices but not the owner's full coat of arms.

Institutional arms from medieval times

An early development was the extension of heraldic design from its use by persons or families to its employment by institutions and associations of various kinds, a consequence of the concept that an assembly or body of people can be personified as an individual, much as a limited company or corporation is viewed as a legal "person." Medieval times provided numerous examples of arms borne by municipalities, churches, and colleges. The arms assumed by an individual or granted to him are regarded as being peculiarly his possession; therefore caution must be used in speaking of family arms. This question can be best dealt with in connection with the royal arms of the sovereign of the United Kingdom.

These arms are borne in their entirety only by the reigning king or queen. No other member of the royal family is permitted to bear the arms without introducing a "difference" mark that will show without doubt that the bearer is not the reigning sovereign. By analogy, the same condition holds for all so-called family arms, which belong to the head of the family; all other members should strictly bear them differenced—that is, with some mark of cadency (a sign indicating the position of the bearer with respect to the head of the family). In Scottish heraldry this rule is very rigidly enforced, but in England and elsewhere it has been allowed to fall into decay, except in the case of the royal family.

Probably the next development in the scope of heraldry was its use by ecclesiastics. The bishops and the abbots of the monasteries used arms on their seals from the 12th century onward. In this variety of heraldic usage, the arms were not those of individuals but of the body they temporarily represented—as also with arms borne by political units such as nations and cities or by educational establishments, many of which date from the Middle Ages. A great extension of medieval heraldry was connected with what came to be called the livery companies. These were guilds or associations of men in trades whose object was to uphold standards of craftsmanship. Most of them obtained charters from the crown and were granted arms. Among numerous examples in Britain are the Grocers, the Mercers, and the Glaziers companies. Membership in these still-existing companies no longer entails practice of their particular trades, but they possess property and have great charitable interests as well as considerable social esteem. Their armorial bearings are of great antiquity and are much displayed on their halls, letterheads, glass, silver, and so forth. Obviously, armorial bearings were assumed in the Middle Ages by such military bodies as the Knights Templars, the Knights of St. John of Jerusalem, the Teutonic Knights, and the great Spanish orders. Military heraldry has continued to the present: each of the three British armed forces, for example, has badges, or in some cases coats of arms, which are in the care of officers of the English College of Arms. The newest of the British armed forces, the Royal Air Force, alone makes use of more than 1,000 coats of arms or badges.

In the 20th century the development of corporate heraldry has gone far beyond anything known before. Throughout the world, banks, insurance companies, and many other great commercial concerns use arms, as do an ever-increasing number of professional, educational, and trade associations.



The arms of U.S. Pres. John F. Kennedy. The gold helmets are a variant on the three silver helmets of an ancient Kennedy coat. The border was added as a further difference. The olive branches and sheaf of arrows are derived from those of the Great Seal of the United States. The lambrequin is blazoned as argent and gules; this is exceptional to the rule that the principal tinctures of the arms (in this case or and sable) be repeated. No motto was included in the grant. The diagonal orientation of the shield is called couché and is optional in all depictions of arms.
Drawing by Wm. A. Norman

Perhaps the event most illuminative of the modern scope of heraldry was the grant, in 1961, by the government of the Republic of Ireland of armorial bearings to the president of the United States of America, John F. Kennedy. Because arms are hereditary and their owners are regarded heraldically as of noble status, the grant amounted to a bestowal of nobility by a state on the head of another state, an occurrence unique in heraldic history.

THE CHIEF COMPONENTS OF ARMORIAL BEARINGS

The trend toward complexity and esoteric significance

Heraldry originated when most men were illiterate but could easily recognize a bold, striking, and simple design. The use of heraldry in medieval warfare enabled combatants to distinguish one mail-clad knight from another and thus to know friend from foe. Thus, simplicity was the principal characteristic of medieval heraldry. In the tournament there was a more elaborate form of heraldic design. When heraldry was no longer used in war and heraldic devices had become a part of civilian life, an intricate type of design evolved with an esoteric significance utterly at variance with its original purpose. In modern times, heraldry has often been looked on as mysterious and a matter for experts only. Indeed, over the centuries its language has become intricate and pedantic. Such intricacy appears ridiculous when it is remembered that in the earlier periods swift recognition of a coat of arms or badge could mean the difference between safety and death, and in many medieval instances battles were lost through a mistake over the sameness of two devices of opposing sides.

The shield. The shield is the essential part of the armorial bearings; without it there can be no heraldic device, except for a woman, a distinction that calls for special treatment and is dealt with later. The word shield can be used to describe the coat of arms but in modern times is seldom employed in this way, except in a poetic context. Armorial bearings are generally referred to more briefly as "arms," or as a "coat of arms," a term derived from the surcoat of silk or linen worn over the armour to keep off the rays of the Sun and to prevent rust from forming on the armour. The surcoat was a waistcoat-like garment on which were shown the same heraldic insignia as on the shield.

Every other object in heraldic achievement is dependent upon the shield or coat of arms. There can be, and quite often is, a coat of arms consisting solely of a shield without any other object, such as a crest surmounted. The arms of the Churchills of Muston, a branch of the same family to which Sir Winston Churchill belonged, have no crest. The reason is that such families possessed arms before crests became fashionable.

The crest. A crest is the object placed on top of the helmet and bound onto it by what is called the wreath of the colours, which shows the two main colours of the shield (see illustration of the royal arms of the sovereign of the United Kingdom). Crests were at first made of leather, later of light wood, and, as time went on, of more valuable materials. It is supposed that they were at first borne in tournaments; they became general in families in England from the 16th century when the venal heralds of that time persuaded crestless families to acquire the addition for a payment. Nowadays, a crest is automatically included in any grant of arms made in England, Scotland, or Ireland.

The crest and the age of the automobile

When horsedrawn carriages were in use, it was the rule to show the whole heraldic insignia on the coach or carriage door. With the advent of motorcars and their smaller door space, the arms were usually left off and only the crest and motto shown. This development may be the reason for the mistake frequently encountered in which the whole coat of arms is referred to as a "crest." It should be emphasized that while a coat of arms can exist without a crest, the existence of a crest without a coat of arms is an impossibility.

The helmet. On top of the shield is placed the helmet, upon which the crest is tied by the wreath. Originally, everything in heraldry was strictly utilitarian. As armorial bearings were used with the suit of armour, there had to be a helmet. In later centuries, rules for the depiction of

Drawing by Wm. A. Norman



The chief components of armorial bearings as indicated on the royal arms of the United Kingdom of Great Britain and Northern Ireland. The royal cipher (ER) is not a part of the arms proper but identifies them as representing Queen Elizabeth II. The roman numeral II is unnecessary here, as the arms of Elizabeth I were different, apart from those of England. The shield shows England (gules three leopards or) quartered with Scotland (or a lion rampant within a double tressure flory counterflory gules) and Ireland (azure a harp or stringed argent). This is the quartering in use since the accession of Queen Victoria in 1837. The shield is encircled by the garter of the Order of the Garter bearing the motto of the order, *Honi soit qui mal y pense*. The dexter supporter, a royally crowned gold lion guardant, and the sinister supporter, a silver unicorn with gold horn, hooves, mane, and tufts and a gold coronet collar and chain, represent England and Scotland, respectively. Atop the full-faced helm of a sovereign with its ermine and gold mantling, or lambrequin, is the royal crown surmounted by the royal crest, a lion statant guardant crowned with the royal crown. The motto *Dieu et mon droit* ("God and my right"), first used by Richard I, appears on the scroll below. The ground beneath the full achievement, called the compartment, is strewn with the floral and plant badges of England (rose), Scotland (thistle), Ireland (shamrock), and Wales (leek).

the helmet were elaborated to show the rank of the bearer, with some displayed in profile and some in full face, and with different metals and accoutrements to indicate status. The shape of the helmet has varied greatly in heraldic representation. While the basic features of heraldry remain unchanged, the modes in which the insignia are shown are subject to change and to fashion. The barrel-shaped helmet was used in the 13th century. The tournament helmet, often shown in drawings, was of a different type altogether, its shape resembling that of a soup tureen.

Mantling. From the helmet hung the mantling or lambrequin. This was of linen or other material and performed the useful function of shielding the wearer from the Sun's rays and also served to catch or deflect sword cuts. The mantling or mantle is painted with the principal colour of the shield, while its lining is of the principal metal. More elaborately styled mantles are used for kings and princes.

Crowns and coronets. These are emblems of the rank of the bearer. With the abolition of most of the great European monarchies, the study of crowns has become mainly of historical and antiquarian interest. The most famous royal crown remaining in the 20th century is that of the United Kingdom; it appears in the sovereign's arms upon the royal helmet and the crest of a golden lion crowned. Coronets (small crowns implying dignity inferior to that of the sovereign) are emblems of rank that are shown, when depicted, between shield and helmet. In Britain there are different coronets specified for the ranks of baron, viscount, earl, marquess, and duke. On the European continent, a much wider use of coronets has prevailed. Among the relics of this usage is the crest coronet, a coronet that supports the crest either instead of the wreath or in addition to it and resting upon it. Another is the chapeau or cap of maintenance, a cap with fur lining that was once worn on the helmet before the development of mantling and that is often used instead of the wreath to support the crest.

The myth of battle cries as sources of mottoes

Mottoes. Many myths have grown up around mottoes. They are often said to have originated as battle cries, but very few actually did. Among those that did are the *Crom a boo,* of the FitzGeralds, the dukes of Leinster, meaning "Crom [one of the family's old castles] forever"; and the "furth fortune and fill the fetters" of some Scottish families, which defies explanation and must refer to some forgotten incident. In succeeding centuries large numbers of mottoes were adopted. They are not part of a coat of arms and can be varied at the user's pleasure, though they are included in a modern grant of arms. More than one motto may be used by the same family. In Scottish arms the motto is usually shown above the crest, in all other countries beneath the arms and always contained in a scroll.

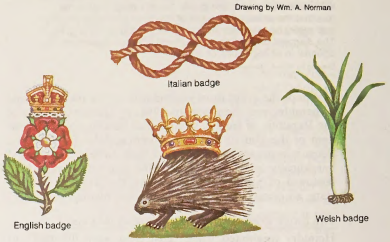
The supporters. These are the figures on either side of the shield of arms and are borne (in English heraldry) by peers, and by other bearers of orders of the highest class, such as Knights of the Garter, the Thistle, St. Patrick, and by Knights Grand Cross. In former times, supporters were used more widely, and a few English families still claim the right. In Scotland their use is much more frequent.

The compartment. The ground or foundation on which the supporters stand is called the compartment. In Scottish arms it is usually a rock or piece of ground strewn with some heraldic object. In England the compartment ought to be shown in the same way, and in the 20th century often is, with the scroll of the motto beneath it; but in the debased heraldic art of the 18th and 19th centuries, the supporters were generally shown as standing on a piece of ironwork or on the scroll.

The achievement. In heraldic writing, the term achievement often carries the same meaning as "arms," but probably its better usage is to describe the whole representation showing shield, helmet, crest, mantling, and supporters. The achievement belongs to only a minority of those who possess arms, since only a few have supporters. In addition, an achievement may include representation of various knightly orders or companionships of knightly orders to which the owner of the arms is entitled. For example, Viscount Montgomery of Alamein, British field marshal and World War II military leader, could show the symbol of the Order of the Garter around his shield; persons with lesser distinctions such as the Distinguished Service Order, Military Cross, and Order of the British Empire may have the decorations shown pendent from their shields. As distinctions of this kind are not hereditary, on the death of the bearer the successor to the arms must not use representations that show these honours.

The badge. This is older than the heraldic system. Such symbols expressing a person, a body, or an impersonal idea are found from ancient times. The eagle of Rome was one of the state's symbols and was the special device of the legions. Many such symbols bring to mind the country they represent; *e.g.,* winged bulls with human faces at once recall Assyria. On Trajan's Column in Rome, devices sometimes bear resemblances to later heraldic designs. On Etruscan vases are seen what not unfairly could be called demi-boars or bulls' heads caboshed. Nearer to heraldic times, the planta genista, or broom plant, which gave its name to the Plantagenet dynasty of England between 1154 and 1485, was a badge of the counts of Anjou before that family had armorial bearings. With the growth of heraldry, badges naturally assumed a heraldic character. They could be varied at the will of the holder, who often had more than one. Badges persist to the present and sometimes accompany a grant of arms.

Badges as symbols of persons or nations



Badges.
English badge: the red rose of Lancaster charged with the white rose of York, surmounted by the royal crown. Italian badge: the knot of the royal house of Savoy. French badge: the porcupine of Orléans, first used by Louis XII; the crown is not always included. Welsh badge: the leek; the daffodil is also a long-established badge of Wales.

Banners and standards. The users of arms in the Middle Ages often displayed them on the fork-tailed pennons of their lances. When the forked ends were cut off, the resulting flag was square, becoming a banner. Particularly valorous conduct was often indicated in this way, and the knight thus distinguished was known as a knight banneret. This last word was sometimes corrupted into baronet, a term that the English antiquary Sir Robert Cotton (1571–1631) used when he suggested to James I of England that he should revive a supposed order of baronets. The resulting Order (1611) was hereditary, however, and had no true connection with medieval knighthood. The banner bears the owner's arms, and in the 20th century anyone who possesses arms is entitled to a heraldic flag in this form. These can sometimes be seen in Great Britain flying over a house. No banner is referred to in the grant of arms made to President Kennedy, but in due course an armorial banner was made for the occasion when his brother, Sen. Robert Kennedy, carried it to the top of a mountain in the Yukon Territory named Mt. Kennedy by the Canadian government in memory of the President. The banner showed the Kennedy arms without the crest. The maker of the banner added a long white streamer on which he put the badge, the latter being part of the Kennedy crest.

On the standard, the main colours of the arms are shown with the owner's badge.

In England flags are often seen flying above churches.

When these show the flag of St. George (England's patron saint), white with a red cross, they carry in the top right-hand corner the arms of the diocese in which the church is situated. Heraldic flags also flew in countries on the Continent in a similar manner. With the disestablishment of heraldic offices in most European countries, contemporary flags are for the most part nonheraldic.

THE ELEMENTS AND GRAMMAR OF HERALDIC DESIGN

Rudiments of heraldry in about 50 terms

Provided that a few elementary principles are grasped, enough knowledge of heraldry can be acquired in a relatively short time to enable the student to understand the meaning of coats of arms. The multitude of terms used in heraldry need not worry him: once the rudiments are learned with some 50 of the terms, the meaning of the large remainder can be ascertained as the occasion arises. For example, when Queen Elizabeth II was crowned, some beautifully carved figures were made of the different badges that had been used by her ancestors, figures now displayed at Hampton Court Palace. They include one very rare badge—a yale. The yale is a mythical heraldic creature. Anyone unfamiliar with it could easily ascertain its meaning from the various heraldic glossaries. It is therefore unnecessary to burden the memory with hundreds of terms (a heraldic glossary generally contains about 800 terms).

The language of heraldry has a curious look. "Azure three wheat sheaves or" has been known to call forth the question, "Or what?" When it is remembered that *or* is the French for gold, the difficulty diminishes. Much heraldic terminology is a quasi-French, archaic language. In the Middle Ages, the French language was used by the ruling class in much of western Europe, so that it was not unnatural that heraldic terms should be French. In England, by about 1400, English words usually were used in preference. Much modern heraldic terminology, however, is so obscure that it seems purposely designed to puzzle the uninitiated.

Drawing by Wm. A. Norman



Heraldic flags.
Banner: the blazon of the shield is applied to the whole surface of a square or a vertically or horizontally oriented rectangular flag. This is the royal banner of Scotland, which follows the blazon of the second quarter of the royal arms of the United Kingdom. Although it is the banner of the sovereign, it is widely but incorrectly used today as the national symbol. Fork-tailed pennon: shown here is that of the Sovereign Military Order of Malta; gules a cross argent. Standard: the cross of St. George at the hoist identifies this as English. The profusion of badges, the diagonally placed motto, and the border of alternating tinctures are typical. This is the standard of Sir Henry Stafford, *c.* 1475.

The terms dexter and sinister mean merely "right" and "left." A shield is understood to be as if held by a user whom the beholder is facing. Thus the side of the shield facing the beholder's left is the dexter or right hand, and that opposite his right, the sinister or left hand.

The field or ground of the shield. The field or ground of the shield is of one of three kinds: a colour, a metal, or a fur. There are five main colours (known as tinctures): azure (blue), gules (red), sable (black), vert (green), and purpure (purple). In English heraldry there are also murrey (sanguine, a tint between gules and purpure), and tenné (an orange-tawny colour). The metals are or (gold) and argent (silver, which is often shown in coloured illustrations as white). The furs are ermine (white field with black spots), ermines (black field with white spots), erminois (gold field with black spots), pean (black field with gold spots), and vair (composed originally of pieces of fur from a species of squirrel that was blue-gray on the back and white underneath, so that when several of its skins were sewn together, the result was a series of cup-shaped figures, alternately blue and white, which is the way vair is shown).

The furs, from ermine to vair

It is considered bad practice to put a colour upon a colour, a metal upon a metal, or a fur upon a fur. Many examples of such bad heraldry are found in old records, but in this as in other instances, the rules now prevalent grew up only very gradually.

The "charges" on the field. The field is said to be "charged" with an object. Heraldic objects are of a large and growing variety; as more and more arms are devised, more and more objects appear as charges—telescopes, aircraft, rolls of newsprint, and so on. Charges are divisible into two classes: the ordinaries (or honourable ordinaries as they are often called, deriving their name from the fact that they are so frequently used) and the others.

The ordinaries. The ordinaries comprise some 20 figures. Among them are: the chief, being a third of the shield and the top part; the pale, a third part of the shield, drawn perpendicularly; the bend, a third part of the shield, drawn from the dexter chief to sinister base; the bend, sinister, drawn from the dexter base to sinister chief; the fess, a third part and taking up the centre of the shield; and the chevron, resembling an inverted stripe in the rank badge of a noncommissioned officer. It should be noted that the bar is a diminutive of the fess, of the same shape, and can be placed in any part of the shield. The term bar sinister is often used in fiction as a synonym for bastardy. It has no such significance, bastardy being denoted heraldically in several other different ways. Since the European nations were Christian when heraldry was invented, the cross appears in many forms in heraldry. The cross in a coat of arms does not imply, however, that the original bearers were crusaders.

The border, or bordure, is used as a mark of difference sometimes, and in English heraldry since the mid-18th century it is used as a sign of bastardy. The orle is an inner border, not touching the ends of the shield; the field is seen within and around the orle, giving it the appearance of a shield with the middle cut out (voided, in heraldry). The tressure, much used in Scottish heraldry, is a diminutive of the orle. The inescutcheon, a small figure shaped like a shield in the middle of the shield, is used to denote the arms of a heraldic heiress; the quarter occupies one-fourth of the shield; the canton, less than the quarter and one-third of the chief; and checky or chequy, the field or charge divided by transverse lines horizontally or perpendicularly into equal parts. Billets are oblong figures. If their number exceeds 10 and they are irregularly placed, the field is described as *billetté*; the paile or pall is the upper half of a saltire (St. Andrew's Cross) and half a pale. The pile is in the shape of an inverted pyramid. The flanch, or flanque, is a segment of a circle drawn from a corner of the chief to the base point. The lozenge is a parallelogram having equal sides forming two acute and two obtuse angles, and a mascle is a lozenge voided. The roundel is circular in form. Lozengy is the field divided

by diagonal lines transversely, and the fret is a mascle interlaced with a saltire.

Minor charges. A field is said to be "powdered" or "semé" when strewn with minor charges; when charged with drops of liquid, it is "gutté." Partition lines divide the shield. The most common ones are straight. "Impalement" or "dimidiation" means the division of the shield into two equal parts by a straight line from the top to bottom. This method is used to show either the arms of husband and wife, the arms of the husband being in the dexter half, or certain types of official arms, as the arms of a bishop's see impaled with his family arms, those of the see being in the dexter half. The shield is divided into four quarters when one coat of arms is quartered with another, as when the children of a heraldic heiress (the eldest daughter of a family of no sons) use their mother's arms with their father's. (Thus in the illustration A = father's arms, B = mother's arms.)



two coats of arms quartered

Other divisions of a shield are: party per pale or simply per pale, division of the field into two equal parts by a perpendicular line (this resembles the impalement just mentioned but does not serve the same purpose of combining arms); party per fess, division into two equal parts by a horizontal line; party per bend; party per chevron; party per saltire; gyronny of eight. When the partition lines are not straight, they can be of several varieties.

The nature and origins of heraldic terminology. Fanciful explanations have been advanced to account for heraldic charges: for example, argent to denote purity, the bend derived from the military cross belt—the cross a sign of a crusading ancestor—and so on. Since no one wrote about heraldry until it had existed for more than 200 years, these explanations of its symbolism can be discounted. With very few exceptions, the origin of the charges is unknown. The Stourton arms ("sable a bend or between six fountains") refer to the six springs in the park of their ancestral estate that are the source of the river Stour. A heraldic fountain does not resemble a real fountain but is put in the form of "roundel wavy argent and azure" (a silver and blue circlet of wavy lines), unless it is expressly stated that the fountain is "proper"; *i.e.,* a natural fountain. The word proper is always used to denote a charge shown in its natural colours or natural form.

The derivation of heraldic charges is more easily discerned in the augmentations of honour, as they are called, when something has been added to a coat of arms by the (British) crown in recognition of services rendered. The arms of the British naval hero Admiral Horatio Nelson show fresh heraldic charges added to his ancestral arms as his victories were gained. Within the past 300 years, augmentations have generally been recorded. An example is the augmentation granted by Queen Victoria to commemorate the discovery by the English explorer John Hanning Speke of the sources of the Nile. The honour, granted posthumously, consisted of the addition to the existing arms of a chief azure upon which appeared a representation of flowing water proper superinscribed with the word Nile in gold lettering. Usually, however, the origins of the various objects used in heraldry are not known. Numerous historical instances of augmentations of honour occurred in continental Europe, especially in connection with the Holy Roman emperors. Frederick II, for example, granted to Conrad Malaspina an augmentation of a chief of the empire, thereby adding an eagle displayed sable to the Malaspina arms of per fess gules and or overall a thorn branch vert with five flowers argent in pale.

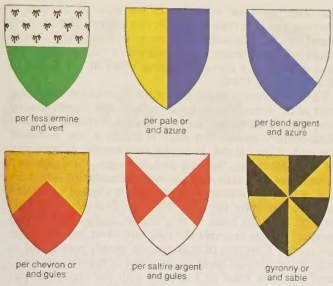
Augmentations of honour

Heraldic descriptions are called blazons. The term is derived from the French *blason,* the etymology of which is uncertain. Originally, it denoted the shield of arms itself and still retains this meaning, but it is now generally used in a derivative sense as meaning the description of the arms. Blazon is thus a noun, and there is also the verb to blazon; *i.e.,* to describe a coat of arms.

There are four generalizations that are useful in the deciphering of blazons. First early coats of arms are simple because they were original and there were so few of them that elaborate differentiation was not required. As time brought many more coats of arms into being, simple coats became much rarer, and the passing of warlike usage made arms much more complicated. Second, punning or canting arms are very common as, for example, trumpets for Trumpington, or a spear for Shakespeare. It is notable, however, that many armorial allusions that were formerly obvious now require research for elucidation. Other allusions have been lost entirely. Third, in grants of arms to people bearing the same name but having no relationship with each other, difference marks have had to be put in.



Ordinaries, basic bearings that may be of any tincture and that may be combined in great variety. A combination of a cross (England) and two saltires (Scotland and Ireland) has resulted in the familiar Union Jack of the United Kingdom. Ermine and certain other textures such as ermines (black with white ermine tails) are regarded as tinctures in their own right and may bear superimposed charges. Discrete charges such as lozenges, mascles, fleurs-de-lis, etc., may be used singly, in pairs, in threes or greater numbers, sometimes in great profusion, as that of ermine tails.



Partition of the shield.
The field is often divided along the lines occupied by ordinaries, just as quartering imitates a cross. "Per fess" means along the line over which a fess would be laid down. The ermine tails illustrated are one type of stylization among many in use. The superior dexter segment on the gyronny shield is called a gyron and is occasionally found singly.
Drawing by Wm. A. Norman

Again, in consequence, blazons have become much more complicated. Finally, in the course of centuries and frequent intermarriages among arms bearers, many quarterly and grandquarterly coats have come about. Quarterly and grandquarterly coats are much more difficult to describe than the simple coats.

Apart from the ordinaries and those other charges that have been mentioned incidentally, there are some peculiarities of heraldic charges that need to be noted. Mythical birds and animals are much used, the product of ancient and medieval natural history—or the lack of it. Such are the dragon, griffin, wyvern, harpy, phoenix, and martlet. In addition, there are some creatures bearing the names of real animals but not resembling them in all respects. The heraldic tiger is more like a lion or a wolf in some features. When the real tiger became known to heraldry, it was described as a Bengal tiger. The heraldic description of animals is very important. Rampant means on the hindlegs, while rampant guardant is the same posture but full faced. Reguardant means looking back; passant, walking. Combattant signifies two animals fighting on hindlegs. Couchant is lying down; dormant, sleeping; and sejant, sitting. A beast of the hunt is called at gaze when looking full face, trippant when at trot with one foot raised, and statant when standing. Parts of an animal may be a charge; *e.g.*, a demilion or demiwolf, a lion's or bear's gamb or foreleg. Heads are described as erased when cut off by a jagged line, couped when cut by a straight line, and caboshed when the severed head looks straight forward. A bird shown with wings expanded is said to be displayed. Creatures placed back to back are addorsed. A fabulous bird, the phoenix, is known to heraldry; also known was the legendary pelican that fed her young on her own blood and was called "in her piety," then being considered an emblem of Jesus Christ, who fed or redeemed his flock with his own blood. The martlet is another fabulous bird widely known outside heraldry, owing to John Milton's reference to the herald's martlet, which has no legs. It is a frequent charge, resembling a swallow, and is used in cadency to denote the fourth son. Other terms have special heraldic significance. Armed is used of the horns, teeth, or claws of a beast, or the beak or talons of a bird, and of the human being when in armour. The term slipped applies to flowers and fruit when the stalk is seen. Counterchanged refers to the field when it has two tinctures, a metal and a colour, and when one is the background for the other on one side of the shield but the relationship is reversed on the other side. An example is the Warner arms: "per bend argent and gules two bendlets between six roses all counterchanged," where the three roses on argent will be gules and the three on the gules will be argent. The field is separated by a partition line and the charge or charges of the arms are said to be countercharged when the charge or portion of a charge that lies on the metal is of the colour and vice versa.

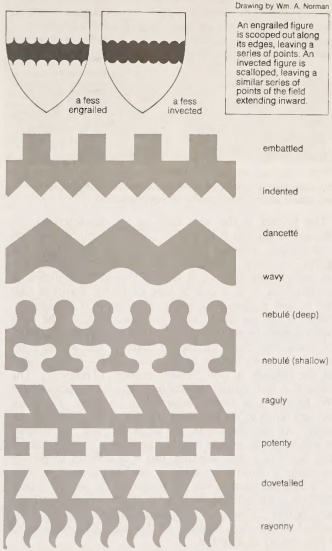
Heraldic descriptions of animals and their postures

THE READING OF HERALDRY

A method has been devised to indicate heraldic colours in black-and-white illustrations. Known as the system of Sylvester Petra-Sancta, an Italian herald, it makes use of the following equivalents: or is denoted by dots or points, azure by horizontal lines, vert by lines from dexter chief to sinister base, purpure by lines from sinister chief to dexter base, argent by a plain field, gules by perpendicular lines, and sable by cross lines horizontal and perpendicular. Furs are depicted with black or white spots on the appropriate ground; vair and countervair are shown by alternate lines and plain surfaces.

The Petra-Sancta system of denoting colour by black-and-white patterns

Describing, or blazoning, of arms must always begin with an identification of the ground of the shield, such as argent or gules or ermine. For a woman who is not married, the arms normally appear on a lozenge, not a shield, but the field or ground in this instance, too, must be the start of the blazon. Then come the charges. A typical blazon is thus: "sable [ground of shield] a chevron ermine between three lions rampant argent crowned or" (arms ascribed to a family of Hinstoke). The chevron is a fur; the lions are silver, appear on the sides of the chevron and its base, and have gold crowns. One important feature in heraldic writing is economy of words. Technically, it should be possible to avoid punctuation marks, thus: "azure a fess between three stags trippant or" (Hind). Here both fess and stags are in gold. When three beasts are depicted, they are shown in the most convenient way around the main charge; that is to say, two in the upper part of the shield



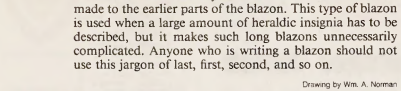
Types of divisions between tinctures.
A line described as flory, or flory counterflory, employs a series of small fleurs-de-lis that have substance of their own beyond the two areas being divided. Rayonny (or *rayonné*) may alternate straight points with curved points.



Canting, or punning, arms, derived from the literal meaning or from the sound of a name.
(Left) Shelley: sable a fess engrailed between three whelk shells or. (Centre) Wellwood: argent an oak tree growing out of a well all proper. (Right) Keyes: per chevron gules and sable, three keys or.
Drawing by Wm. A. Norman

and one below. A straightforward coat with only one charge on the field is that of the Italian Segni family of Agnani, which gave to the church the popes Innocent III, Gregory IX, and Alexander IV: "gules an eagle displayed chequy sable and or." Economy, however, can be carried too far as in the following: "azure a lion rampant double queued barry of ten argent and gules armed and langued of the last crowned or, within a bordure of the second and third" (Mountbatten). Here is an example of a usage that grew up in past centuries and was designed to avoid repetition of the name of a tincture but in reality served to create confusion. "Of the last" means that the lion's claws and tongue are in red or gules. "Of the second and third" means simply argent and gules. There is no real economy since more words actually are used and reference has been made to the earlier parts of the blazon. This type of blazon is used when a large amount of heraldic insignia has to be described, but it makes such long blazons unnecessarily complicated. Anyone who is writing a blazon should not use this jargon of last, first, second, and so on.

Drawing by Wm. A. Norman



Conventional representations of tinctures used by engravers and others when actual colours are not practicable.

The helmet is the next item to be characterized, although in blazons it is usually taken for granted and left undescribed. When it is mentioned, it is said to be "befitting his degree." Although the helmet need not appear in written descriptions, it always should be depicted in illustrations. It is a bad feature of many drawings that the helmet is absent, showing the crest as if it were airborne above the shield and thus unsupported. The crest must always be mentioned when, as in the vast majority of cases, one exists. In formal blazons, the wreath (also called the torse) is given as well; thus, crest—"on a wreath of the colours, a wolf passant proper" (Trelawny). The wreath is not usually mentioned, however, because like the helmet it is always assumed to be there. The term colours refers to the two chief colours of the ground of the shield. As with the shield, the older the crest, the simpler it will be. Most people can envisage on a knight's helmet the figure of a wolf walking, but it is difficult to picture someone in armour wearing as his head crest "the stern of a Spanish man-of-war on waves of the sea all proper thereon inscribed 'San Josef' with the motto above, 'Faith and Works' " (Nelson). This latter example belongs to the period of decadent heraldry in the late 18th century and 19th century in England.

Blazoning the crest

The mantling, or lambrequin, is mentioned in formal descriptions but not in general usage. The supporters and compartment pertain only to a few classes of arms bearers, and in descriptions the supporters are blazoned after the crest (or crests). The compartment is not usually described but sometimes has to be, as in the arms of the earl of Perth: supporters (two savages = two ancient Caledonians) stand on "a compartment strewn with caltraps" (from "caltrops," iron instruments designed to maim horses' feet and used by the Scots with great effect at the Battle of Bannockburn, 1314).

The motto comes at the end of the description, not being part of the arms. The badge is rarely found, except among very historic families (and by a strange inversion in some 20th-century grants), but when it occurs it, too, comes at the end of the blazon. It can be very simple, as with that of Lord Mowbray, Segrave, and Stourton—"a sledge or." It may be very elaborate, as with Constantine—"a hurt [*i.e.*, a roundel in azure] charged with a leopard's face and surmounted upon the edge with two fleurs-de-lis in pale or, and as many roses in fess, argent, barbed and seeded proper." In this example the roses are silver but the leaves are proper. Coronets of rank are not usually mentioned in English or Scottish heraldry, but caps of maintenance and crest coronets must be blazoned with the crest. Banners and standards are not as a rule mentioned in blazons, though they may be when they occur in a modern grant.

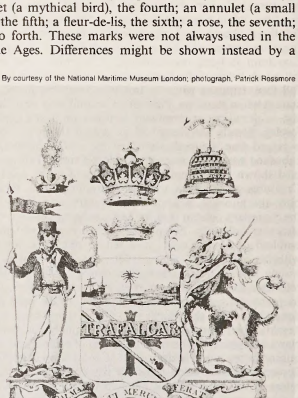
MANIPULATION OF HERALDIC DESIGN

It is clear that the vast majority of heraldic charges are without the foundation of legend often assigned to them, though in many instances the real origin of the charge is lost. A Jacobean dramatist could write of a family as old as the first virtue that merited an escutcheon—a nice poetic flourish that should remain in poetic realms. In modern grants the heralds try to give some allusion to the grantee's work and achievements and his place of origin, and canting arms still appear. Many arms are recondite, however, in their significance, and much has to be learned before the significance of the charges is known.

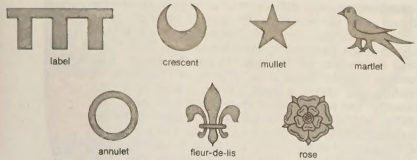
The use of devices to show position in family

Cadency. The rules evolved over the centuries to denote particular distinctions in heraldry are fairly straightforward. Cadency is the use of various devices designed to show a man's position in a family, with the aforementioned basic aim of reserving the entire arms to the head of the family and to differentiate the arms of the rest, who are the cadets, or younger members. Heraldic works in the 16th century refer to cadency marks as: a label for the eldest son during his father's lifetime; a crescent for the second son; a mullet (five-pointed star) for the third; a martlet (a mythical bird), the fourth; an annulet (a small ring), the fifth; a fleur-de-lis, the sixth; a rose, the seventh; and so forth. These marks were not always used in the Middle Ages. Differences might be shown instead by a

By courtesy of the National Maritime Museum London; photograph, Patrick Rossmore



Earl Nelson's coat of arms, drawn in sepia, 1806.



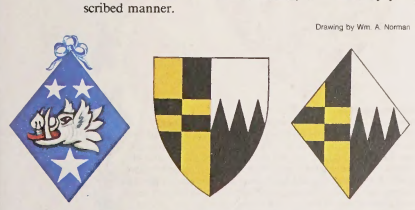
Marks of cadency, used to difference the arms of cadets of the same family. The label, the mark of the eldest male heir, is a notable feature of the arms of the prince of Wales, the heir to the throne The second cadet displays a crescent, the third a mullet, and so on. These symbols may be of any tincture and may be used otherwise than as signs of cadency.

Drawing by Wm. A. Norman

change of tincture, by adding small charges to the field, and the like. Both on the Continent and in England, rules of cadency have long ceased to be used. It is customary for all members of a family to use the entire arms of the head. There are, however, two exceptions. Very occasionally, a crescent is used for difference by a noble family showing descent from a second son. The other exception occurs in the arms of the British royal family, in which the cadency system exists in rigour. The reason is that the royal arms are arms of sovereignty and cannot be shared. The sovereign alone can have the whole undifferenced arms. Nor does any member of the royal family—not even the prince of Wales—have any right to the use of arms until they have been granted to him by the sovereign. A label of difference with marks is placed on the arms; a three-pronged label for the children of the sovereign, a five-pronged label for grandchildren. The Duke of Windsor, after his abdication as Edward VIII in 1936, was granted arms with a label.

In Scotland the position on cadency is very different. Since heraldry is regulated in Scotland by acts of the Scots Parliament before the Union in 1707 with England, and confirmed by the British Parliament, the regulation of arms is very precise. The strict observance of cadency is probably because the Celtic clans formed the original social system in Scotland before the advent of feudalism. Thus only the chief of the name can have the entire arms. He matriculates, or enters his arms, in the registers of the Lord Lyon King of Arms (whose court has jurisdiction over armorial bearings in Scotland). This registration also applies to his eldest son (subject to suitable differencing of the arms) who inherits them in due course. The younger sons must petition for a matriculation of the paternal arms with a suitable difference indicating the position of each in the family. As families from the descendants of the original grantee continue to be established, so there is matriculation and rematriculation, in a carefully prescribed manner.

Drawing by Wm. A. Norman



unmarried woman married woman widow

Arms of women.
A woman adopts the undifferenced arms of her father. An unmarried woman: the arms of her father are displayed on a lozenge; a true lover's knot signifies unmarried status. A married woman: the wife's arms, to the sinister, impale those of the husband, to the dexter; the husband displays the combined arms as head of the family, and the wife shares his escutcheon. A widow: the woman's arms revert to the lozenge, retaining the deceased husband's impaled arms.

Arms of women. Arms of women are shown during spinsterhood or widowhood on a lozenge, not on a shield, without a crest (except in Scotland, where a woman who is chief of a clan and head of the name, such as MacLeod of MacLeod, is allowed a crest). A woman divorced and not remarried also uses a lozenge. The arms of a married woman are shown in conjunction with her husband's by impalement, the division of the shield into two equal portions, the husband's arms on the dexter and the wife's on the sinister. Should she be a heraldic heiress, the arms of her family are placed upon an inescutcheon (originally inner escutcheon) or "escutcheon of pretence" (a small shield whose position in the fess point of the husband's shield gives it precedence over all of the other parts of the shield). The only exception to these rules for women is that of a queen regnant like Elizabeth II, who, being sovereign and thus considered in heraldic terms to be herself the source of honour for all her subjects, possesses the full arms of sovereignty of her royal house and kingdom.

In the Middle Ages, arms for women were often shown on shields, and those like Joan of Arc who bore arms in battle may have used crests.

Picturepoin



Royal arms of England as used from Henry V's time, quarterly France and England, first used by Edward III but by him with the French quarter showing strewn fleurs-de-lis. Roof boss from St. George's Chapel, Windsor Castle, 1483–1528.

Quarterings and marshalling. In the quarterings and the marshalling arrangement of more than one coat of arms in the same shield, the position of heiresses must be considered first. The children of a heraldic heiress are entitled on her death to quarter her arms with their father's (the arrangement is to show the shield divided into four quarters so that quarters 1 and 4 are the father's arms, 2 and 3 the mother's). This positioning of the quarterings is also used in England when an additional surname and arms are taken, almost always in obedience to a will. This is the "name and arms" clause peculiar to English law. Its operation over the last 200 years is responsible for the double- or treble-barrelled surname common in England and also found in Scotland. Thus in Salusbury-Trelawny, the original Trelawny arms appear in quarters 1 and 4 and the assumed additional arms for Salusbury in 2 and 3. A famous historical case is that of King Edward III of England, who in 1340 claimed the throne of France in right of his mother, a French princess. He then quartered the lilies of France (the fleurs-de-lis) with the lions (leopards) of England. England should have been placed in 1 and 4, but Edward gave this position to France, probably because of the greater size and resources of France at that time. In this form, the royal arms continued until 1800, when the empty title of king of France was dropped and the lilies went out with it.

When quarterings are inherited from a woman, no crest is transmitted with them, because a woman cannot pass



Heirship through the female lines.
Drawing by Wm. A. Norman

on a crest. The matter alters, however, when additional arms are taken in obedience to a will; then a double crest is likely. There is no reason why a further assumption may not occur, so that triple or quadruple hyphenated names are found: for example, the English county family Sawbridge-Erle-Drax has quarterly arms, 1 and 4 Drax, 2 Erle, and 3 Sawbridge. This type of quartering is not difficult to follow, but a real problem in marshalling several coats in one shield arises when more than one heraldic heiress occurs in the same family. Some families of long descent have often married heraldic heiresses, acquiring many quarterings. Sometimes several hundred quarterings are attributed to the head of a great family. A splendid instance of quartering occurred in the achievement of the empress Maria Theresa. Before her accession to the imperial throne she was queen of Hungary and Bohemia and by marriage grand duchess of Tuscany. As a sovereign in her own right she bore a shield on which there were 29 quarterings. The dukes of Rohan-Chabot in France bore a quartered shield with Navarre, Scotland, Brittany, and Flanders; overall was an escutcheon of Rohan quartering Chabot.

Problem of marshalling several coats in one shield

Even without such large numbers of arms to deal with, the marshalling of quarterings is still a problem. Various methods have been tried, often with results as difficult to decipher. An interesting illustration of the marshalling of several coats of arms is that of the baronet Cameron-Ramsay-Fairfax-Lucy. The arms are said to be quarterly with the arms of Lucy in 1 and 4. Then in 2, the description runs, grand quarter counterquartered. This means that quarter 2 is itself a quarterly coat, 1 and 4 of which are for Fairfax, 2 for Ramsay, with the third quarter counterquartered; *i.e.*, itself quartered, showing two coats of arms. The third quarter is that for Cameron.

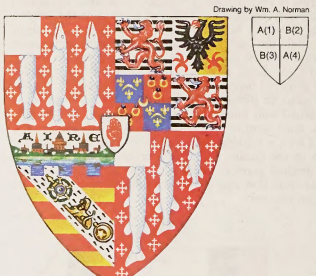
Arms of bastardy. The heraldic illustration of bastardy was achieved in several ways in the older days of chivalry. Little social or moral obloquy attended the status of bastardy, possibly because of the late marriages of the upper classes and their arranged unions. Thus the arms of a bastard were merely differentiated as perhaps those of a distant cadet line would be. From early times the bend or bendlet sinister was used. The erroneous use of bar sinister might have come about because the French for bend sinister was *une barre*. In the arms of British royal bastards like the dukes of St. Albans, of Buccleuch, of Grafton, and others, a baton sinister is used to denote bastardy. The royal arms of these illegitimate scions are said to be "debruised" (crossed or partly covered) by the baton sinister. English heralds in the last 150 years have often signified bastardy by the use of the bordure.

Nonfamilial heraldry. The arms of grantees other than families or individuals are often encountered. In modern times the granting of arms to private persons has ceased in some countries where grants of corporate arms are frequent. In Sweden, for example, grant of neither title nor arms occurs, but grant of arms to public bodies, such as local administrative units, is frequent. Historically, it was an easy passage from the arms of individuals to those of corporate bodies. This is particularly evident in the military sphere, where the great crusading orders led to the many important orders of chivalry in the principal European countries. The Elephant of Denmark, the Golden Fleece of Spain and of Austria, the Holy Spirit of France, the Garter of England, and the Thistle of Scotland were all preceded by the orders of military monks, and all have insignia that contain heraldic features and that occur in many arms illustrations. Most of the older bishops' sees have official arms; in the Anglican Church the missionary bishops as well as the diocesan bishops have arms. In the Roman Catholic Church the episcopal sees all have arms; and new arms are granted by the pope, who, as head of the Vatican state, is a temporal sovereign as well as spiritual head of the church. The arms of the popes often contain charges that are added to their individual arms after their election to the papacy or to earlier ecclesiastical office.

The military orders of chivalry

Dominion and colonial arms are necessarily interconnected with royal arms, since the British crown in each country is or was the source of honour and must have granted arms to its various territories. Because of the vast extent of the former British Empire, the richest collection

Drawing by Wm. A. Norman



Marshalling of several coats of arms.
(Left) The arms of the Cameron-Ramsay-Fairfax-Lucy family, blazoned: arms-quarterly, 1st and 4th gules semé of cross-crosslets, three lucies hauriant argent, a canton of the last (Lucy); 2nd, grand quarter counterquartered, 1st and 4th argent, three bars gemel sable surmounted of a lion rampant gules, armed and langued azure (Fairfax); 2nd parted per pale argent and or, an eagle displayed sable, armed beaked and membered gules (Ramsay); 3rd counter-quartered, 1st and 4th azure a branch of palm between three fleurs-de-lis or; 2nd and 3rd gules three annulets or stoned azure. In the centre of these quarters a crescent or (Montgomerie); 3rd gules, three bars or, on a bend ermine, a sphinx between the badge of the royal (Portuguese Order of the Tower and Sword) and the gold medal presented to Col. John Cameron of Fassifern by command of the Grand Signior, in testimony of that sovereign's high sense of his services in Egypt, and on a chief embattled a representation of the town of Aire in France, all proper (Cameron of Fassifern).
(Right) Representation of two coats of arms quartered.



Marks of bastardy.
These are common marks of illegitimacy but do not invariably have that meaning. (Left) The arms of the duke of St. Albans debruised by a baton sinister, in this case charged with three roses. (Right, top) The bordure wavy (or a bordure wavy sable). (Centre) The bordure compony (vert a bordure compony argent and gules). (Bottom) The baton sinister (purpure a baton sinister argent).
Drawing by Wm. A. Norman

of arms of dominion is to be found in the numerous members of the Commonwealth. Canada, for instance, has arms for the sovereign of Canada as authorized by George V in 1921, and the 12 provinces of Canada have similar arms approved by the sovereign. Much the same is true of the former French colonies, though there was no sovereign to grant the arms. The arms of political units are used throughout the Western world. The cities and boroughs of the United Kingdom, for example, have their heraldry, as do the U.S. states.

The blazoning of these nonfamilial arms is conducted on the same principles as for family arms, except that the explanation of the charges is usually forthcoming.

Drawing by Wm. A. Norman



The arms of Canada, derived from the royal arms. The three maple leaves in the base, originally green, were altered to red to conform to the new national flag adopted in 1964. French traditions and ties are represented by the gold fleurs-de-lis and the white lilies in the compartment.

Historical development of heraldry

The exact date or place of origin of the heraldic system in western Europe in the 12th century is not known; neither are the precise reasons for its introduction. But limits can be drawn to indicate generally when heraldry began, and probable reasons for its emergence can be deduced. The Bayeux Tapestry, produced in the last quarter of the 11th century in Bayeux, France, is a pictorial record that shows conditions in everyday life and in warfare at the time of the Norman Conquest of England. The English and Norman soldiers are armed alike. None of the Englishmen has a design of any kind on his shield or armour. In a few scenes, Normans or Frenchmen have designs on their shields that have a rough heraldic resemblance. In scene VIII of the tapestry, four of the followers of Guy, count of Ponthieu, have shields with these devices: some kind of creature holding what appears to be a fish in its mouth, a rough design emerging from the left side of the shield, a cross, and an animal resembling a sheep. In scene XII the messenger of William the Conqueror bears a winged creature on his shield, and this reappears in scenes XIII and XV. In scene XVIII a cross or a variant of it is seen on a shield, but this was probably a boss (metal protuberance) to strengthen it. Scene LXXV has a Norman knight with a design of a birdlike creature on the shield, but generally the Normans shields have only bosses in the middle.

Bayeux Tapestry and its record of shields

In 1066, at the time of the Conquest, heraldry clearly did not exist. The most that can be said is that possibly some of the rudiments out of which it emerged were present. Even at the time of the First Crusade (1095–99), there is evidence that heraldry was not yet in use. The *Alexiad,* the history of the Emperor Alexius (reigned 1081–1118) written by his daughter, the princess Anna Comnena, contains a vivid description of the Frankish barbarians—as the Crusaders appeared in the eyes of the civilized Byzantines. The Princess gave a careful account of the Crusaders' armour. She said that Alexius exhorted his archers to shoot at the Franks' horses rather than their riders, whose armour rendered them almost invulnerable. "For the Frankish weapon of defence is this coat of mail, ring plaited into ring, and the iron fabric is such excellent iron that it repels arrows and keeps the wearer's skin unhurt. An additional weapon of defence is a shield which is not round, but a long shield, very broad at the top and running out to a point, hollowed out slightly inside, but externally smooth and gleaming with a brilliant boss of molten brass." The shields in her description, therefore, are similar to those depicted in the Bayeux Tapestry. Thus, all the pictures from later times of King Alfred the Great, William the Conqueror, or Charlemagne bearing coat armour can be disregarded as anachronisms.

EARLY ROOTS OF HERALDRY

Not until a generation after the First Crusade does unmistakable evidence of heraldic designs appear. The earliest evidence is in an enamel (Musée de Tessé, Le Mans, France) made not later than 1151 showing Geoffrey, count of Anjou, bearing a shield azure with possibly four rampant golden lions (the exact number is not discernible because of the position in which the shield was depicted). The count was the son-in-law of King Henry I of England (reigned 1100–35), and, according to a chronicle, Henry in knighting Geoffrey bestowed upon him a shield that bore painted lions. In addition, from 1136, heraldic devices appear on seals. It is possible that the insignia were used first on seals and later on warriors' shields. Simultaneously, body armour was becoming all-enveloping, and some means of distinguishing men in full armour became necessary. The heavy barrel-type helmet closed in the wearer's face except for the opening of his visor, and a mixture of plate and ring mail enclosed the whole body. Another factor was the Crusades, in which men from different lands had to be distinguished from one another.

The first unmistakable evidence of heraldic designs

Within a few years heraldry was found throughout all of western Christendom. The first English king to bear arms was the crusader Richard I the Lion-Heart (1157–99). The three gold leopards or lions of England have been used by every dynasty since his time.



Plaque from the tomb of Geoffrey Plantagenet, count of Anjou, enamel, Limoges school, *c.* 1151–60. The stylized pattern of blue and white lining the figure's cloak represents a series of squirrel skins, called vair, frequently mentioned in blazons. In the Musée de Tessé, Le Mans, France.

Giraudon—Art Resource/EB Inc

Seals. The earliest body of evidence of heraldic insignia is found in seals, large numbers of which have been preserved in England, France, and Germany, with fewer surviving in Spain and Italy. For the first century of heraldry, seals supply the bulk of information. It is from seals that the rise and development of the English royal arms can be traced. Seals from the first years of Richard I's reign show the design of a lion rampant to the left side. Some scholars think that two lions were used, since only half of the shield can be seen. Seals from the end of Richard's reign bear the three lions or leopards that have been used by all subsequent English sovereigns. The adoption of the same coat of arms by subsequent dynasties is also found in the royal arms of Sweden and of Denmark; but unlike the English, those royal families place their family arms on an inescutcheon in the middle of the shield.

Roll of arms. Next to seals as evidence of heraldic usage come the rolls of arms, which in England date from about 1250. These are lists of arms, often with pictures drawn ("tricked") on the rolls, of persons who were present on a particular occasion, such as at a tournament or on a military expedition. England and Belgium (Flanders) are rich in the rolls of arms. France, Spain, and Scotland have fewer surviving examples. In place of the rolls, collections of painted books of arms have been preserved in Germany. A notable roll is the *Armorial de Berry*, dating about 1445, the work of a French herald, Gilles le Bouvier, who travelled widely and recorded arms borne in France, England, Scotland, Germany, Italy, and other European countries.

The *Armorial*

Records in stone and glass. Another very important source of information is to be found in representations on stone, wood, glass, and in books and engravings. Over the gateway of Bodiam Castle in Sussex can be seen the arms cut in stone of three owners of the castle, the families of Bodiam (who took their name from the place), Wardedieux, and Dalyngrygge. Of such arms nothing would be known without these centuries-old memorials. In Rome, many examples occur of the arms of various popes in their palaces and other buildings, for instance the bees of the Barbarini pope Urban VIII in the Palazzo Venezia. Heraldic glass is usually much more recent in origin but of immense value in supplying information as it is always in colour, while other memorials often are not. Very few churches of any great age in western Europe are without armorial illustration. Switzerland in particular has splendid memorials in stained glass; for example, the Dom, or main Protestant church in Berne, has windows that are aflame with glorious heraldic colours. Sweden has a fine collection of coloured plaques of arms in the House of Nobles in Stockholm; in the Frederiksborg Slot (castle) at Hillerød, Denmark, may be seen the shields of the Knights of the Order of the Elephant, in which can be read the history of heraldry over several centuries.

Church brasses. Brasses in churches are an important source of heraldic information. It was formerly the custom to put a brass tablet over the grave slab, and on this would be shown a figure of the deceased with his armorial bearings. Many fine examples of this are found in old English churches. A very fine collection of floor brasses is in the small church of Stopham in Sussex, which has been the memorial place of the local Barttellot family for many centuries. Also found in churches are hatchments, heraldic paintings on wood that were made for deceased persons and hung over their house doors, being later set up in the local church where they have often been preserved.

By courtesy of the Society of Antiquaries, London



Page from the *Armorial de Berry*, by Gilles le Bouvier *c.* 1445, showing the simplicity of the early coats. In the collection of the Society of Antiquaries, London.



Sir Nicholas Hawberk, rubbing from his tomb brass, Cobham Church, Kent, 1407. Hawberk's arms appear alone on the dexter and impaled with the arms of his wife on the sinister.
By courtesy of Robert Lind

As for written material—such as official enactments, grants of arms, and books—nothing is dated earlier than the 14th century.

GROWTH OF HERALDRY AFTER THE 13TH CENTURY

The initial meaning of the term herald is uncertain. Some authorities derive the word from two German words—*Heer,* "a host," and *Held,* "a champion"—not a very obvious etymology. It is clear that heralds were in existence from the 13th century, if not earlier. In their beginnings they were more or less menials, ranked with the jugglers and the minstrels. First used as messengers who wore livery, they later were used in the jousts and tourneys to announce the contenders. To identify the knights, they had to know the arms on the shields, and from this grew their knowledge and skill in heraldry.

Messengers and aides at jousts and tourneys

Heraldic colleges and offices. From this lowly origin have come the colourful figures of the English College of Arms, who now alone, save for the Scottish heralds, possess a high position in the modern world. The Lord Lyon, the head of the Scottish heralds, derives his office from a much higher source than do the heralds in other parts of Europe. The Sennachie, or official bard of the Scots' king, was the record keeper of the old Celtic kingdom of Scotland, and from the Sennachie is derived the Lord Lyon, a great officer of state in Scotland.

The older statements found in many books that the medieval heralds were either identical, or in some way connected, with the old Greek *kēryx* or Latin *fetialis* need only be stated to be dismissed. Since ancient times men have been found who, because their persons were accepted as sacred, were able to carry messages and other communications between nations either hostile or strange to one another. These ambassadors bore several names before the development of a diplomatic corps. In the earlier Middle Ages, for instance, churchmen, monks, or priests were used for this type of service. When William I the Conqueror sent a messenger to Harold II of England, it was a monk who carried William's denunciation of Harold. Heralds were not then in existence.

As they ascended the social scale, heralds began to serve as ambassadors between the different courts, a function that was still theirs in the first half of the 17th century. In 1627, for example, Sir Henry St. George was joined in a commission with Lord Spencer and Peter Young to present the insignia of the Order of the Garter to Gustavus II Adolphus, king of Sweden, who then knighted Sir Henry and granted him an augmentation to his arms showing the royal arms of Sweden.

At first every great noble had his herald, and the royal heralds were distinguished from the others by the greater importance of their masters. Gradually, it came about that a king would form his heralds into a college or corporation. The King of France did so in 1407; it was not until 1484 that the King of England followed by establishing the College of Arms, which has been housed for 300 years in the same building in London. The English College, sometimes called the Heralds' College, has outlived all similar elaborate establishments in Europe, except that in Scotland. Outside Great Britain, heraldic offices are found in the 20th century in Sweden, Denmark, the Republic of Ireland, and Spain.

The English College

The English College is under the control of the earl marshal, an office that has been hereditary in the family of the duke of Norfolk since the 1660s. The holder of the dukedom is always earl marshal. Under him are 13 officers of arms; three kings of arms (Garter, Norroy and Ulster, and Clarenceux); six heralds (Windsor, Richmond, York, Lancaster, Chester, and Somerset); and four pursuivants (Rouge Dragon, Rouge Croix, Bluemantle, and Portcullis). These medieval names are derived from sources connected with royalty, titles, badges, or orders of knighthood. Pursuivants are "followers," or junior heralds. In Scotland the Lord Lyon is the head of the heraldic officers, of whom there are three heralds (Albany, Marchmont, and Rothesay) and four pursuivants (Carrick, Kintyre, Unicorn, and Ormond). In England and Scotland the officers are not civil servants but members of the sovereign's household. In both countries there are also heralds extraordinary, who are appointed at times for special reasons or functions.

Heraldic Visitations from 1530 to 1686

Records and grants. In England an important development came with the Heraldic Visitations. From 1530 in the reign of Henry VIII to 1686 in the reign of James II, commissions were issued by the sovereign to the heralds directing them to proceed to a county in England or Wales and to inspect the arms in use there. The records of the Visitations have been preserved and constitute a valuable body of genealogy as well as of heraldry. From the period of the Visitations the heralds built up huge collections of family history and began to record pedigrees.

From about a half century before the foundation of the College of Arms, the English heralds are found to be issuing grants of arms on behalf of the sovereign. This is some 300 years after the first appearance of heraldry, which obviously much antedated not only royal colleges or corporations of heralds but even the existence of heralds themselves. From this evidence, it seems clear that in the early days of heraldry men assumed arms to suit themselves without reference to any authority. A very simple coat of arms would not be difficult to invent. That three unrelated persons from three different counties could bear these same arms is not only not surprising but proof that the arms were self-chosen. When disputes over ownership among the three came up, the matter was referred to the king. His judgment was final, but it is noteworthy that one of the defeated, compelled to give up his arms, then consoled himself with a new coat that was also self-derived and self-assumed. Unquestionably, the great majority of ancient coats of arms, borne before 1500, were never granted but were taken by the owners.



Detail of the "Grant of arms to the Worshipful Company of Tallow Chandlers," London, 1456. The herald in the illuminated initial wears a tabard that bears the king's arms front, back, and repeated on the sleeves, and a coronet as a mark of royal authority. In the collection of the Worshipful Company of Tallow Chandlers, London.

By courtesy of the Court of the Worshipful Company of Tallow Chandlers, London

Writers on heraldry. The earliest writing on heraldry extant is a short treatise by Bartolo da Sassoferrato, whose *Tractatus de insigniis et armis* ("Treatise on Insignia and Arms") was published about 1356. In his small book Bartolo describes the various categories of arms bearing and how they have been assumed. He refers specifically to arms granted by a prince and gives reasons for their value but asks why one man may not bear arms identical with those of another.

In 1355 Bartolo had been sent to Pisa from Perugia as an envoy to the Holy Roman emperor, Charles IV, from whom he reccived many privileges, including a grant of arms, which were the same as those of the Emperor as king of Bohemia but with changed tincture: "or a lion rampant with two tails gules." An American scholar, L.M. Mladen, remarked of this grant and others made by Charles IV at the same time: "Charles was in all probability the first ruler ever to grant arms. To my knowledge, no earlier occurrence has been found."

The first English heraldic writer was John of Guildford, or Johannes de Bado Aureo, whose *Tractatus de armis* ("Treatise on Arms") was produced about 1394. Then came a Welsh treatise, the *Llyfr arfau* ("Book of Arms"). Nicholas Upton, a canon of Salisbury Cathedral, about 1440 wrote *De studio militari* ("On Military Studies"). John of Guildford's treatise was printed in 1654 with Upton's work and the *Aspilogia* of Sir Henry Spelman by Sir Edward Bysshe, Garter king of arms, who edited and annotated all three works. The whole was in Latin; no complete English version of Upton's book has been published.

The trend to the nonsensical after the 15th century

These books are by authorities who were concerned with the realities of heraldry in their own day. A tendency away from actuality and toward the fanciful and absurd manifested itself from the end of the 15th century. Some of these farfetched conceits showed themselves in *The Boke of St. Albans* (1486). Yet by comparison with the vast mass of nonsense contained in the folios of the 16th century, those conceits were reasonable. The works of Sir John Ferne, *Blazon of Gentrie* (1586), Gerard Legh, *The Accedens of Armorie* (1562), and John Guillim, *A Display of Heraldrie* (1610) not only perpetuate the nonsensical natural history of olden days but are largely responsible for the belief that heraldic charges have a definite symbolic meaning and that they were granted as the reward of valorous deeds.

Continental versus British heraldry. Much greater significance was attached in former times to heraldic insignia, though the attitude varied from country to country. Heraldry has become more widespread than at any other time, but as a sign of rank it has hardly any remaining value.

A distinction can be made between the Continent and Great Britain regarding medieval and later heraldry. The doctrine of the Seize Quartiers (16 quarterings) prevailed over most of the Continent but not in Britain. This theory required that, in order for a person to claim nobility, all of his 16 ancestors (16 great-great-grandparents) should have been entitled to bear arms. This is known as the "Proof of the Seize Quartiers" and was the reason why Frederick II the Great of Prussia, though professing the views of the Enlightenment in the Age of Reason, diligently scrutinized his courtiers' quarterings. The theory is based on the rigidity of a noble caste that married only with its own kind. On the Continent, every member of a noble family is noble; hence the enormous numbers of titles. Similarly, the continental royalty tended to marry only with other royal families. As a result both royal and noble families formed a class apart from the bulk of the people.

Seize Quartiers

Continental heraldic insignia, therefore, from their origins until the late 18th century, provided symbols to indicate a higher caste and, in fact, were signs of nobility. Yet, strangely, in several countries heraldry was in wide general use as a means of identification, serving in the same way as a surname. In France, for example, it is abundantly clear that from the 13th century, not only the bourgeoisie of the towns but also the peasants bore heraldic arms. The usage had percolated down from the noble class. The earliest example of the use of arms by a peasant is that of Jaquier le Brebiet in 1369, whose arms show a punning allusion to his name (*brebis*, "sheep"—three sheep held by a girl). In other European lands—Hungary and the Low Countries—burgher or peasant arms were also found, but neither in these lands nor in France were the possessors regarded as noble.

Peasant and bourgeois arms

In France the regulations of arms followed a quite different course from those in England. Although King Charles VI had in 1407 led the way in creating a college of arms, his heralds lost influence over the next two centuries. They had no power, unlike their Scottish and English counterparts, to grant arms and they gradually faded into insignificance. To overcome the loss, Louis XIII in 1615 appointed a *juge général d'armes* ("general judge of arms"), an official whose powers resembled those of the Lord Lyon. The French royal government showed itself very broad-minded on the possession of arms. A decree of Louis XIV in 1696, designed to raise money, ordered all persons who bore arms to register them. Even those who were not part of the arms-bearing population were forced to buy arms. Later, in 1760, an ordinance was framed by which the lesser townsfolk, artisans, and peasants were to be excluded from the use of arms—after these classes had used them for 400 years. But the Parlement of Paris refused to allow the ordinance to be implemented. Later, during the French Revolution (1789), arms were suppressed as signs of feudalism.

The view of arms held in England was and is quite different. No such thing as a noble caste has ever existed in England. Only the reigning peer and his wife are regarded as noble; the rest of the family are commoners, and only a few of them bear what are called courtesy titles. Moreover, except for the Hanoverian (1714–1837) and Victorian (1837–1901) epochs, the royal house has not necessarily married royalty but much more often the nobility, a practice to which it has returned in the present century. As a result, the Continental doctrine of Seize Quartiers does

not apply in England. It cannot, since noble and non-noble are so mixed. Nor are there any such things as non-noble arms. All arms are on the same basis; all are signs of gentility—nobility, in fact. Arms then have long had a high social significance in England; those who possess them have social prestige. The situation is somewhat different in Scotland where arms are bound up with clanship and family solidarity. To the Scotsman, arms are not so much a matter of social status as of family or clan, and he proves his right to them by process of law in Lyon Court.

A matter of clan and family in Scotland

20TH-CENTURY HERALDRY

France and Italy. Without a monarchy, heraldry can still flourish, but it does not normally do so. The French Revolution abolished arms, which returned with the monarchy, and now, although France is a republic, a person assuming arms to which he is not entitled may be prosecuted. In the same way, it is not permissible to assume a name of a great family. In England, however, the assumption of the name Windsor (the name of the British royal family) is possible for anyone. There is, however, no recognized heraldic authority in France, nor for that matter in other European countries that have abolished their monarchies. Most such republican states have associations that seek to maintain heraldic standards. Thus in Italy the Collegio Araldico (Heraldic College) consists of experts whose main object is to promote heraldic and genealogical studies. An association of nobles was formed as the National Heraldic Council of the Italian nobility, under the authority of former king Umberto II; it tries to regulate the use of arms and titles.

Many changes in heraldry have taken place since 1945, as it is not a static subject.

Communist countries. In communist European countries, the study of genealogy and heraldry had been generally suppressed. Since about 1956 it had been possible to obtain much more statistical information from the U.S.S.R. (until its dissolution in 1991) than formerly, but no heraldic data were supplied. Much the same was true of the other communist countries in Europe. Under communist regimes, heraldry was viewed as it was by the French revolutionaries, as part of the feudal past, although there is reason to believe that heraldic archives were in many cases carefully preserved.

Ireland. A different development occurred in two of the republics that have emerged from the British Empire—Ireland and South Africa—both of which have set up their own heraldic offices. As early as 1382, there was an Ireland King of Arms who was responsible for all matters armorial in that country. The last holder of the office died in 1487, and in 1553 Edward VI created a new armorial king under the title of Ulster, to control bearings throughout Ireland. His place of business was in Dublin Castle. When the Irish Free State, or Eire (now the Republic of Ireland), was established in 1921–22, the Ulster Office was reserved as an appointment of the British crown with the then-current Ulster to hold office for life. After his death in 1940, an arrangement was made between the British and Irish governments by which the heraldic office in Dublin Castle with its records was taken over by the Irish authorities. Photostat copies were made of the records and sent to the College of Arms, London. The Irish government appointed a Chief Herald of Ireland, and the Ulster Office became known as the Genealogical Office. A civil servant was then appointed as Chief Herald of Ireland. The office of Ulster King has now been united with that of Norroy King in the College of Arms, London. The Irish Herald carries out the duties formerly performed by Ulster in the 26 counties of the Republic of Ireland; Norroy and Ulster has jurisdiction over the six counties of Northern Ireland (Ulster).

Chief Herald of Ireland

South Africa. In South Africa an act was passed in 1962 under which was established a Bureau of Heraldry and a Heraldry Council for the grants, registration, and protection of coats of arms, badges, and other emblems. A state herald is appointed as head of the Bureau of Heraldry. The Heraldry Council consists of the state herald and at least seven other members appointed by the government minister responsible.

The United States. There has been a remarkable evolution of heraldry in the United States. Ever since the American Revolution, the use of arms, especially of arms of English families with whom the users were related or whose surname they bore, has continued. The English College of Arms claims heraldic jurisdiction over persons of English and Welsh descent (Wales has been reckoned with England in this and all other administrative matters since the union of England and Wales, 1542). The Lord Lyon of Scotland claims jurisdiction likewise over persons of Scottish descent throughout the world. In addition, the English College at one time claimed a worldwide imperial jurisdiction over anyone who could be brought within the definition of British subject. Under this jurisdiction even the Indian princes were occasionally granted arms by the College of Arms, although they were not British subjects but independent rulers who had entered into treaty relations with the British crown. Many Americans have been granted arms by the college by virtue of their descent from English or Welsh forebears or by the Lord Lyon if they are of Scottish descent. Irish Americans often were granted arms from Dublin, from either the Ulster King or his successor. Americans of Northern Irish descent have been granted arms by Norroy and Ulster. In addition, there are several states of the United States that were formerly Spanish territory, and the Spanish Kings of Arms, the equivalent of the English and Scottish heralds, exercised a heraldic authority over persons of Spanish descent in the old Spanish Empire. By extension, they have recently granted arms to Americans who are resident in those formerly Spanish states but who are not of Spanish descent.

Heraldry in states from the old Spanish Empire

To these classes of arms obtained by Americans from overseas must be added such instances as the award of arms to Pres. Dwight D. Eisenhower in Denmark. Also, Americans of French, German, Italian, Polish, and other European origins have inherited from their immigrant ancestors arms once granted or recorded by heraldic authorities no longer in existence. All these classes of arms share one feature whatever their origin: they are hereditary honours granted to American citizens by other countries. As they do not carry titles, they do not contravene the principle of the American Constitution on this subject. An American who receives a knighthood of some foreign state possesses only an honorary knighthood; he is not "Sir." But for the citizen of an independent sovereign power to approach and receive from another power a hereditary honour has seemed to many Americans an undesirable procedure. There have been and still are thousands of assumptions of arms by Americans on consideration of mere mail-order salesmanship; arms in these transactions are, at the best, supposed to be those "of the name" (that is, belonging to a name rather than a person), a view for which no justification exists.

Endeavours have been made to set up American authorities who would not only record but also grant arms. The New England Historic Genealogical Society of Boston appointed a Committee on Heraldry that since 1928 has issued rolls of arms, in which have been entered the names and arms of those who have submitted their claims to its judgment. The use of this method of issuing or publicizing arms recalls the usage previously referred to, which has not been practiced in Europe for more than 400 years. In the introduction to the second roll (1932) it is stated: "There is certainly no legal reason, perhaps no reason at all, why an American gentleman should not assume *in more majorum* any new coat that pleases his fancy, but he should not assume an old coat, for if he does, he is very likely denying his own forefathers and he surely is affirming what he has no sufficient reason to be true." Not only British-derived arms but also continental European arms have been registered. In addition, the committee has assisted inquirers in devising new coats of arms, not only for schools, colleges, and other institutions but also for individuals. In the introduction to the first roll a very reasonable view toward heraldry was expressed: "Taking into consideration the early history of coat armour there seems to be no reason in this country at least, why anyone provided he observes the simple rules of blazon and does not appropriate the arms of another, may not assume and use

any coat he desires." The American College of Heraldry and Arms, Inc., was established in the state of Maryland in 1966. It has two divisions: The American College of Arms, which is concerned with the arms of individuals, their registration, and, more importantly, the granting of arms; and the College of Arms of the United States, which deals with such items as arms, crests, and standards for corporate concerns. Arms were granted to Pres. Lyndon B. Johnson during his term of office, to Pres. Richard M. Nixon, and to Vice Pres. Spiro T. Agnew.

Revival of England's ancient Court of Chivalry

England. Heraldic development has also occurred in England, where in 1954 the ancient Court of Chivalry was revived. This was once the court of the lord high constable and the earl marshal, and it dealt with matters relating to knights and gentlemen. Although it was concerned also with matters of military discipline, it was not the forerunner of the modern court-martial in the armed forces. The court gradually declined in the 17th and 18th centuries and had not sat from 1735 until its revival.

The office of lord high constable has long ceased to be hereditary or of permanent status in England. During coronations, however, a constable is appointed for the occasion. Therefore, in the revived court that sat in 1954 to deal with a test case, *Manchester Corporation* v. *Manchester Palace of Varieties,* the earl marshal (the duke of Norfolk) presided with a surrogate, who was the lord chief justice of England appearing in his capacity as a doctor of civil law. As a result of the sitting, the jurisdiction of the court was confirmed, and the City of London recorded its crest and supporters.

The unorganized condition of heraldry in many European countries has spurred private attempts to bring some order into the field. The movement known as the International Congresses of Heraldry and Genealogy began in 1928 with a meeting in Barcelona, Spain. A second Congress was held in Rome and Naples in 1953, and from that time regular meetings occurred at two- or three-year intervals. Out of these was established an international organization, El Instituto Internacional de Genealogia y Heraldica (The International Institute of Genealogy and Heraldry), with its headquarters in Madrid.

USES OF HERALDRY FOR STUDY AND VERIFICATION

The uses of heraldry, apart from its general significance in providing distinguished symbols, are considerable. Heraldry illustrates much history and literature that is otherwise obscure. Heraldry on buildings, in manuscripts, and in paintings is of immense value for purposes of identification. It serves to link one person with another, to connect families, and to disclose origins of states and of institutions. With the use of heraldry in connection with genealogy, from which it cannot easily be separated, much that is difficult to follow becomes easy to interpret. In every building that contains armorial engravings or other pictures of arms, there is a concise history of the place, which gives a knowledgeable onlooker a useful clue to the background and details of the building.

In law the place of heraldry is sometimes very precise, as in modern Scotland; in England, on the other hand, it is more involved. To understand the latter is to gain an insight into the development of English law. Similarity of arms does not always indicate identity of family; it may merely denote assumption of devices. According to the English laws of inheritance, not only an estate but also a surname and arms can pass by eventual succession to individuals unconnected by blood with the original owner. Thus in the families of Lytton and Carew, and in branches of Trelawny, instances occur in which possession of the name and arms is contrary to blood relationship. Sometimes, however, identity of a name with a puzzling discrepancy in the arms can only be explained by a study of the family's heraldry.

BIBLIOGRAPHY. There are many guides to heraldry; the majority appear to be derived from an old original. Among the most useful for the beginner are: A.C. FOX-DAVIES, *Heraldry Explained* (1906, reprinted 1971), a small, well-illustrated book written by the clearest expositor of heraldry; CHARLES MACKINNON, *The Observer's Book of Heraldry* (1966), a very useful and clearly written work; and SIR ANTHONY R. WAGNER, *Heraldry in England* (1946, reprinted 1953), a brief account but with 15 plates in colour; LESLIE G. PINE, *Teach Yourself Heraldry and Genealogy,* 2nd ed. (1970), designed for the beginner and including a glossary of terms, and *The Genealogist's Encyclopaedia* (1969, reprinted 1977); and JAMES R. PLANCHÉ, *The Pursuivant of Arms or Heraldry Founded upon Facts* (1852, reprinted 1973).

More specialized works include A.C. FOX-DAVIES, *A Complete Guide to Heraldry* (1909, reprinted 1978), and *Armorial Families: A Directory of Gentlemen of Coat Armour,* 7th ed., 2 vol. (1929–30, reprinted 1970). GERARD J. BRAULT, *Early Blazon: Heraldic Terminology in the Twelfth and Thirteenth Centuries, with Special Reference to Arthurian Literature* (1972), illustrates the development of the language of blazon. The writings of OSWALD BARRON should be read as a correction to some of Fox-Davies' theories on heraldic history. Some of his best work may be found in the 12 volumes of *The Ancestor,* which he edited in 1902–05. Two volumes now reprinted in one constitute *A Treatise on Heraldry: British and Foreign* by JOHN WOODWARD and GEORGE BURNETT (1892, reprinted 1969), which contains some inaccuracies. SIR THOMAS INNES, *Scots Heraldry,* new ed. (1978), explains much in heraldic practice that may not otherwise be clear. Another good view of heraldry in the international sense is ROBERT GAYRE OF GAYRE AND NIGG, *The Nature of Arms* (1961). For information on Irish heraldry, see SIR CHRISTOPHER and ADRIAN LYNCH-ROBINSON, *Intelligible Heraldry* (1948, reprinted 1967). *Boutell's Heraldry* has been edited many times since its first appearance in the 19th century. The editions of V. WHEELER-HOLOHAN (1931), C.W. SCOTT-GILES (1958), and JOHN P. BROOKE-LITTLE (1978) are all very useful, as are JOHN P. BROOKE-LITTLE, *An Heraldic Alphabet* (1973); and OTTFRIED NEUBECKER and JOHN P. BROOKE-LITTLE, *Heraldry: Sources, Symbols, and Meaning* (1977). JULIAN FRANKLYN, *Shield and Crest,* 3rd ed. (1971), gives much otherwise not easily obtained information.

Little continental work on heraldry has been translated into English. Perhaps the best course is to consult periodical publications like *The Augustan,* with articles from a wide variety of sources. Several non-English-language encyclopaedias, such as the *Enciclopedia Italiana* (1929–37), have excellent articles on heraldry. Many non-English writers transcend their national boundaries in writing on the subject. RÉMI MATHIEU in *Le Système héraldique français* (1946) writes of French heraldry and helps to clarify the whole subject. On Spanish heraldry, see JOSÉ ASENSIO Y TORRES, *Tratado de heráldica y blasón,* 3rd ed. rev. (1854, reprinted 1929); and LUCAS DE PALACIO, *De Genealogia y heraldica* (1946), the latter author being especially interested in the connection between totemism and heraldry; the Japanese *mon* is exhaustively dealt with by CARROLL PARISH in *The Augustan,* vol. 11, no. 1 (1968).

Other works that will help the student after he has acquired a sound knowledge of this subject are: SIR ANTHONY R. WAGNER, *Heralds and Heraldry in the Middle Ages,* 2nd ed. (1956, reprinted 1960), and *Historic Heraldry of Britain* (1939, reprinted 1972); and LESLIE G. PINE, *The Story of Heraldry* (1952, reprinted 1967), an account of much controversial matter, including the present English heraldic position in law; GEORGE D. SQUIBB, *The High Court of Chivalry: A Study of the Civil Law in England* (1959); C. PAMA, *Lions and Virgins* (1965), which discusses the history of arms in South Africa; and LESLIE G. PINE, *International Heraldry* (1970), and *American Origins* (1960, reprinted 1980), concerned with heraldry throughout the world.

Reference works in English are easily available. SIR JOHN B. BURKE, *The General Armory of England, Scotland, Ireland, and Wales* (1883, reprinted 1976), gives the description under alphabetical order of surnames of thousands of coats of arms. JOHN WOODY PAPWORTH, *An Alphabetical Dictionary of Coats of Arms Belonging to Families in Great Britain and Ireland* (1858–74, reprinted 1977), the counterpart of Burke's *General Armory,* enables the seeker to trace a coat of arms without knowing the owner's name. Both books contain inaccuracies. JAMES PARKER, *A Glossary of Terms Used in Heraldry,* new ed. (1970), with 1,000 illustrations, is a helpful book, as is LESLIE G. PINE, *A Dictionary of Mottoes* (1983). Regular editions of JOHN DEBRETT, *Peerage, Baronetage, Knightage and Companionage* (since 1713); and JOHN B. BURKE, *A Genealogical and Heraldic History of the Peerage, Baronetage, and Knightage* (since 1826), and *Genealogical and Heraldic History of the Landed Gentry* (since 1837), abound in illustrations and descriptions of arms. The role of the herald in history and in the contemporary art and science of heraldry is explored in RODNEY DENNYS, *The Heraldic Imagination* (1976), and *Heraldry and the Herald* (1982).

(L.G.P.)

Hinduism

The term Hinduism refers to the civilization of the Hindus (originally, the inhabitants of the land of the Indus River). Introduced in about 1830 by British writers, it properly denotes the Indian civilization of approximately the last 2,000 years, which evolved from Vedism, the religion of the Indo-European peoples who settled in India in the last centuries of the 2nd millennium BC.

Because it integrates a variety of elements, Hinduism constitutes a complex but largely continuous whole and has religious, social, economic, literary, and artistic aspects. As a religion, Hinduism is a composite of diverse doctrines, cults, and ways of life.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 823, and the *Index*. This article is divided into the following sections:

General nature and characteristic features 519
  Common characteristics of Hindu belief 519
  Three margas: paths to salvation 520
The history of Hinduism 521
  Sources of Hinduism 521
  The prehistoric period (3rd and 2nd millennia BC) 522
  The Vedic period (2nd millennium–7th century BC) 522
  Challenges to Brahmanism (7th–2nd century BC) 522
  Early Hinduism (2nd century BC–4th century AD) 523
  The rise of devotional Hinduism (4th–11th century) 524
  Hinduism under Islām (11th–19th century) 525
  The modern period (19th–20th century) 527
Sacred texts 529
  Vedas 529
  Sutras, Shastras, and Smritis 533
  Epics and Purāṇas 534
  Vaishnavism and Śaivism 537
  Philosophical texts 540
  Tantrism 542
  Vernacular literatures 545
  Folk Hinduism 547
Rituals, social practices, and institutions 549
  Sacrifice and worship 549
  Sacred times and places 551
  Ritual and social status 552
Cultural expressions: visual arts, theatre, and dance 554
  Types of symbols 554
  The arts 554
The place of Hinduism in world religions 555
  Hinduism and other religions of Indian origin 555
  Hinduism and Islām 556
  Hinduism and Christianity 556
Bibliography 557

General nature and characteristic features

The spectrum that ranges from the level of popular Hindu belief to that of elaborate ritual technique and philosophical speculation is very broad and is attended by many stages of transition and varieties of coexistence. Magic rites, animal worship, and belief in demons are often combined with the worship of more or less personal gods or with mysticism, asceticism, and abstract and profound theological systems or esoteric doctrines. The worship of local deities does not exclude the belief in pan-Indian higher gods or even in a single high God. Such local deities are also frequently looked upon as manifestations of a high God.

Religious pluralism

In principle, Hinduism incorporates all forms of belief and worship without necessitating the selection or elimination of any. It is axiomatic that no religious idea in India ever dies or is superseded—it is merely combined with the new ideas that arise in response to it. Hindus are inclined to revere the divine in every manifestation, whatever it may be, and are doctrinally tolerant, allowing others—including both Hindus and non-Hindus—whatever beliefs suit them best. A Hindu may embrace a non-Hindu religion without ceasing to be a Hindu, and because Hindus are disposed to think synthetically and to regard other forms of worship, strange gods, and divergent doctrines as inadequate rather than wrong or objectionable, they tend to believe that the highest divine powers complement one another. Few religious ideas are considered to be irreconcilable. The core of religion does not depend on the existence or nonexistence of God or on whether there is one god or many. Because religious truth is said to transcend all verbal definition, it is not conceived in dogmatic terms. Moreover, the tendency of Hindus to distinguish themselves from others on the basis of practice (orthopraxy) rather than doctrine (orthodoxy) further de-emphasizes doctrinal differences.

Hinduism is both a civilization and a congregation of religions; it has neither a beginning or founder, nor a central authority, hierarchy, or organization. Every attempt at a specific definition of Hinduism has proved unsatisfactory in one way or another, the more so because the finest scholars of Hinduism, including Hindus themselves, have emphasized different aspects of the whole.

COMMON CHARACTERISTICS OF HINDU BELIEF

Nevertheless, it is possible to discern among the myriad forms of Hinduism several common characteristics of belief and practice.

Authority of the Veda and the Brahman class. Perhaps the defining characteristic of Hindu belief is the recognition of the Veda, the most ancient body of religious literature, as an absolute authority revealing fundamental and unassailable truth. At the same time, however, its content has long been practically unknown to most Hindus, and it is seldom drawn upon for literal information or advice. Still, it is venerated from a distance by every traditional Hindu, and those Indians who reject its authority (such as Buddhists and Jains) are regarded as unfaithful to their tradition. The Veda is also regarded as the basis of all the later Shastraic texts used in Hindu doctrine and practice. Parts of the Veda are still quoted in essential Hindu rituals, and it is the source of many enduring patterns of Hindu thought.

Also characteristic of Hinduism is the belief in the power of the Brahmans, a priestly class possessing spiritual supremacy by birth. As special manifestations of religious power and as bearers and teachers of the Veda, Brahmans are considered to represent the ideal of ritual purity and social prestige.

Doctrine of atman-brahman. Hindus believe in an uncreated, eternal, infinite, transcendent, and all-embracing principle, which, "comprising in itself being and non-being," is the sole reality, the ultimate cause and foundation, source, and goal of all existence. This ultimate reality is called *brahman*. As the All, *brahman* causes the universe and all beings to emanate from itself, transforms itself into the universe, or assumes its appearance. *Brahman* is in all things and is the Self (atman) of all living beings. *Brahman* is the creator, preserver, or transformer

Identity of individual self and the ultimate reality

and reabsorber of everything. Although it is Being in itself, without attributes and qualities and hence impersonal, it may also be conceived of as a personal high God, usually as Vishnu (Viṣṇu) or Śiva. This fundamental belief in and the essentially religious search for ultimate reality—*i.e.*, the One that is the All—have continued almost unaltered for more than 30 centuries and have been the central focus of India's spiritual life.

Ahimsa: non-injury. A further characteristic of Hinduism is the ideal of ahimsa. Ahimsa, "non-injury" or the absence of the desire to harm, is regarded by Indian thinkers as one of the keystones of their ethics. Historically, ahimsa is unrelated to vegetarianism; in ancient India, killing people in war or in capital punishment and killing animals in Vedic sacrifices were acceptable to many people who for other reasons refrained from eating meat. However, the two movements, ahimsa and vegetarianism, reinforced one another through the common concept of the disinclination to kill and eat animals, and together they contributed to the growing importance of the protection and veneration of the cow, which gives food without having to be killed. Neither ahimsa nor vegetarianism ever found full acceptance. Even today, many Hindus eat beef, and nonviolence (as the ideal of ahimsa is often translated) has never been a notable characteristic of Hindu behaviour.

Doctrines of transmigration and karma. Hindus generally accept the doctrine of transmigration and rebirth and the complementary belief in karma, or previous acts as the factor that determines the condition into which a being, after a stay in heaven or hell, is reborn in one form or another. The whole process of rebirths is called samsara. Any earthly process is viewed as cyclic, and all worldly existence is subject to the cycle. Samsara has no beginning and, in most cases, no end; it is not a cycle of progress or a process of purification but a matter of perpetual attachment. Karma, acting like a clockwork that, while running down, always winds itself up, binds the atmans (selves) of beings to the world and compels them to go through an endless series of births and deaths. This belief is indissolubly connected with the traditional Indian views of society and earthly life, and any social interaction (particularly those involving sex or food) results in the mutual exchange of good and bad karma. It has given rise to the belief that any misfortune is the effect of karma, or one's own deeds, and to the conviction that the course of world history is conditioned by collective karma.

Samsara and moksha (liberation)

Such doctrines encourage the view that mundane life is not true existence and that human endeavour should be directed toward a permanent interruption of the mechanism of karma and transmigration—that is, toward final emancipation (moksha), toward escaping forever from the impermanence that is an inescapable feature of mundane existence. In this view the only goal is the one permanent and eternal principle: the One, God, *brahman,* which is totally opposite to any phenomenal existence. Anyone who has not fully realized that his being is identical with *brahman* is thus seen as deluded. The only possible solution consists in the realization that the kernel of human personality (atman) really is *brahman* and that it is their attachment to worldly objects that prevents people from reaching salvation and eternal peace. (Hindus sometimes use the largely Buddhist term nirvana to describe this state.)

Concepts of iṣṭadevatā and Trimūrti. Although those Hindus who particularly worship either Vishnu or Śiva generally consider one or the other as their "favourite god" (*iṣṭadevatā*) and as the Lord (Īśāna) and Brahman in its personal aspect, Vishnu is often regarded as a special manifestation of the preservative aspect of the Supreme and Śiva as that of the destructive function. Another deity, Brahmā, the creator, remains in the background as a demiurge. These three great figures (Brahmā, Vishnu, and Śiva) constitute the so-called Hindu Trinity (Trimūrti, "the One or Whole with Three Forms"). This conception attempts to synthesize and harmonize the conviction that the Supreme Power is singular with the plurality of gods in daily religious worship. Although the concept of the Trimūrti assigns a position of special importance to some great gods, it never has become a living element in the religion of the people. Moreover, Brahmā has had no major cult since ancient times, and many Hindus worship neither Śiva nor Vishnu but one or more of the innumerable other Hindu gods.

Ashramas: the four stages of life. In the West, the so-called life-negating aspects of Hinduism have often been overemphasized. The polarity of asceticism and sensuality, which assumed the form of a conflict between the aspiration to liberation and the heartfelt desire to have descendants and continue earthly life, manifested itself in Hindu social life as the tension between the different goals and stages of life. The relative value of an active life and the performance of meritorious works (*pravṛtti*) as opposed to the renunciation of all worldly interests and activity (*nivṛtti*) was a much-debated issue. While one-sided religious and philosophical works, such as the Upanishads, placed emphasis on renunciation, the dharma texts argued that the householder who maintains his sacred fire, procreates children, and performs his ritual duties well also earns religious merit. Nearly 2,000 years ago, these dharma texts elaborated the social doctrine of the four ashramas (stages of life). This concept is an attempt at harmonizing the conflicting tendencies of Hinduism into one system. It held that a member of the three higher classes should first become a chaste student (brahmachari); then become a married householder (grihastha), discharging his debts to his ancestors by begetting sons and to the gods by sacrificing; then retire (as a vanaprastha), with or without his wife, to the forest to devote himself to spiritual contemplation; and finally, but not mandatorily, become a homeless wandering ascetic (sannyasin). The situation of the forest dweller was always a delicate compromise that remained problematic on the mythological level and was often omitted or rejected in practical life.

The householder

Although the status of a householder was often extolled, and some authorities, regarding studentship as a mere preparation, went so far as to brand the other stages as inferior, there were always people who became wandering ascetics immediately after studentship. Theorists were inclined to reconcile the divergent views and practices by allowing the ascetic way of life to those who are, owing to the effects of restrained conduct in former lives, entirely free from worldly desire, even if they had not gone through the traditional prior stages.

THREE MARGAS: PATHS TO SALVATION

Hindus disagree about the way (marga) to final emancipation (moksha). Three paths to salvation (variously valued but nonexclusive) are presented in an extremely influential religious text, the *Bhagavadgītā* ("Song of the Lord"; *c.* 200 BC), according to which it is not acts themselves but the desire for their results that produces karma and thus attachment. These three ways to salvation are (1) the karma-marga ("the path of duties"), the disinterested discharge of ritual and social obligations; (2) the jnana-marga ("the path of knowledge"), the use of meditative concentration preceded by a long and systematic ethical and contemplative training, yoga, to gain a supra-intellectual insight into one's identity with *brahman;* and (3) the bhakti-marga ("the path of devotion"), the devotion to a personal God. These ways are regarded as suited to various types of people.

Although the search for moksha has never been the goal of more than a small minority of Hindus, liberation was a religious ideal that affected all lives. Moksha determined not only the hierarchical values of Indian social institutions and religious doctrines and practices but also the function of Indian philosophy, which is to discuss what one must do to find true fulfillment and what one has to realize, by direct experience, in order to escape from samsara (bondage) and obtain spiritual freedom. While those who have not been reached by formal Indian philosophy have only vague ideas about the doctrines of karma and moksha, in semipopular milieus these doctrines gave rise to much speculation.

For the ordinary Hindu, the main aim of worldly life lies in conforming to social and ritual duties, to the traditional rules of conduct for one's caste, family, and profession.

Dharma: comprehensive system of duties, morals, and religious law

Such requirements constitute an individual's dharma (law and duties), one's own part of the broader stability, law, order, and fundamental equilibrium in the cosmos, nature, and society. *Sanātana* (traditional) dharma—a term used by Hindus to denote their own religion—is a close approximation to "religious practices" in the West. This traditional dharma applies theoretically to all Hindus, but it is superseded by the more particular dharmas that are appropriate to each of the four major varnas, or classes of society: Brahmans (priests), Kṣatriyas (warrior kings), Vaiśyas (the common people), and Śūdras (servants). These four rather abstract categories are further superseded by the more practically applicable dharmas appropriate to each of the thousands of particular castes (*jātis*). Thus, religion for Hindus is mainly a tradition and a heritage, a way of life and a mode of thought. In practice, it is the right application of methods for securing both welfare in this life and a good condition in the hereafter. (W.Do.)

The history of Hinduism

The history of Hinduism began in India about 1500 BC. Although its literature can be traced only to before 1000 BC, evidence of Hinduism's earlier antecedents is derived from archaeology, comparative philology, and comparative religion.

SOURCES OF HINDUISM

The religion of the Rigveda

Indo-European sources. The earliest literary source for the history of Hinduism is the Rigveda (Ṛgveda), the hymns of which were chiefly composed during the last two or three centuries of the 2nd millennium BC. The religious life reflected in this text is not that of Hinduism but of an earlier sacrificial religious system, generally known as Brahmanism or Vedism, which developed in India among Aryan invaders. This branch of a related group of nomadic and seminomadic tribal peoples originally inhabiting the steppe country of southern Russia and Central Asia brought with them the horse and chariot and the Sanskrit language. Other branches of these peoples penetrated into Europe, bringing with them Indo-European languages that developed into the chief language groups now spoken there.

Before they entered the Indian subcontinent (*c.* 1500 BC), the Aryans were in close contact with the ancestors of the Iranians, as evidenced by similarities between Sanskrit and the earliest surviving Iranian languages. Thus, the religion of the Rigveda contains elements from three evolutionary strata: an early element common to most of the Indo-European tribes; a later element held in common with the early Iranians; and an element acquired in the Indian subcontinent itself, after the main Aryan migrations. Hinduism arose from the continued accretion of further elements derived from the original non-Aryan inhabitants, from outside sources, and from the geniuses of individual reformers at all periods.

Hinduism has a few direct survivals from its Indo-European heritage. Some of the rituals of the Hindu wedding ceremony, notably the circumambulation of the sacred fire and the cult of the domestic fire itself, have their roots in the remote Indo-European past. The same is probably true of the custom of cremation and some aspects of the ancestor cult. The Rigveda contains many other Indo-European elements, such as the worship of male sky gods with sacrifices and the existence of the old sky god Dyaus, whose name is cognate with those of the classical Zeus of Greece and Jupiter of Rome ("Father Jove"). The Vedic heaven, the "world of the fathers," resembled the Germanic Valhalla and seems also to be an Indo-European inheritance.

Cremation and sacrifice

Indo-Iranian sources. The Indo-Iranian element in later Hinduism is chiefly found in the initiatory ceremony (*upanayana*) performed by boys of the three upper classes, a rite both in Hinduism and in Zoroastrianism that involves the tying of a sacred cord. The Vedic god Varuṇa, now an unimportant sea god, appears in the Rigveda as sharing many features of the Zoroastrian Ahura Mazdā ("Wise Lord"); the hallucinogenic sacred drink soma corresponds to the sacred *haoma* of Zoroastrianism.

Indigenous sources. Even in the earlier parts of the Rigveda the religion had already acquired numerous specifically Indian features. Some of the chief gods, for example, have no clear Indo-European or Indo-Iranian counterparts. Although some of the new features may have evolved entirely within the Aryan framework, it is generally presumed that many of them stem from the influence of the indigenous inhabitants. The Vedic Aryans may never have been in direct contact with the civilization of the Indus Valley in its prime, but the religion of the valley's culture undoubtedly influenced them.

Non-Indo-European sources. *The Dravidian hypothesis.* Features of Hinduism that cannot be traced to the Rigveda are sometimes ascribed to the influence of the original inhabitants, who are often vaguely and incorrectly referred to as "Dravidians." The ruling classes of the Harappā culture (*c.* 2500–1700 BC), or the Indus civilization, may have spoken a Dravidian language, but as long as their script remains undeciphered this cannot be proved. Moreover, the presence of Dravidian speakers throughout the whole subcontinent at any time in history is not attested. The Mediterranean racial type, to which most modern higher-caste Dravidian speakers belong, is widespread throughout India; but it cannot be proved that all people of this type originally spoke Dravidian languages or that all followed the same culture. Equally or more widely spread in South and Southeast Asia is the Proto-Australoid racial type, the purest members of which in India are the tribal peoples of the centre and the south, many of whom speak languages of the Austric family. Thus, although many aspects of Hinduism are traceable to non-Aryan influence, not all of these aspects are borrowed from "Dravidians." In the 20th century the term Dravidian generally refers to a family of languages and not to an ethnic group.

Dravidian culture

Other sources. The Central Asian nomads who entered India in the two centuries before and after the beginning of the Christian Era might have influenced the growth of devotional Hinduism out of Vedic religion. The classical Western world directly affected Hindu religious art, and several features of Hinduism can be traced to Zoroastrianism. The influence of later Chinese Taoism on Tantric Hinduism (an esoteric system of rituals for spiritual power) has been suggested, though not proved. In more recent centuries, the influence of Islām and Christianity on Hinduism can be seen.

The process of "Sanskritization." The development of Hinduism can be interpreted as a constant interaction between the religion of the upper social groups, represented by the Brahmans (priests and teachers), and the religion of other groups. From the time of the Aryan invasion (*c.* 1500 BC) the indigenous inhabitants of the subcontinent have tended to adapt their religious and social life to Brahmanic norms. This has developed from the desire of lower-class groups to rise on the social ladder by adopting the ways and beliefs of the higher castes. This process, sometimes called "Sanskritization," began in Vedic times when non-Aryan chieftains accepted the ministrations of Brahmans and thus achieved social status for themselves and their subjects. It was probably the principal method by which Hinduism spread through the subcontinent and into Southeast Asia. Sanskritization still continues in the form of the conversion of tribal groups, and it is reflected by the persistent tendency of low-caste Hindus to try to raise their status by adopting high-caste customs, such as wearing the sacred cord and becoming vegetarians.

If Sanskritization has been the main means of spreading Hinduism throughout the subcontinent, its converse process, which has no convenient label, has been one of the means whereby Hinduism has changed and developed over the centuries. The Aryan conquerors lived side by side with the indigenous inhabitants of the subcontinent, and many features of Hinduism, as distinct from Vedic religion, may have been adapted from the religions of the non-Aryan peoples of India. The phallic emblem of the god Śiva arose from a combination of the phallic aspects of the Vedic god Indra and a non-Vedic icon of early popular fertility cults. Many features of Hindu mythology and several of the lesser gods—such as Gaṇeśa, an elephant-headed god, and Hanumān, the monkey god—were

Contributions from non-Aryan religions

incorporated into Hinduism and assimilated into the appropriate Vedic gods by this means. Similarly, the worship of many goddesses who are now regarded as the consorts of the great male Hindu gods, as well as the worship of the one great goddess herself, may have originally incorporated the worship of non-Aryan local goddesses. Unorthodox circles on the fringes of Brahmanic culture (probably in southern India) were one of the important sources of the system of ecstatic devotional religion known as bhakti.

Thus, the history of Hinduism can be interpreted as the imposition of orthodox custom upon wider and wider ranges of people and, complementarily, as the survival of features of non-Aryan religions that gained strength steadily until they were adapted by the Brahmans.

THE PREHISTORIC PERIOD (3RD AND 2ND MILLENNIA BC)

Indigenous prehistoric religion. The prehistoric culture of the Indus Valley arose in the latter centuries of the 3rd millennium BC from the metal-using village cultures of the region. There is considerable evidence of the religious life of the Indus people, but until their writing is deciphered its interpretation is speculative. Enough evidence exists, however, to show that several features of later Hinduism had prehistoric origins.

In most of the village cultures, small terra-cotta figurines of women, found in large quantities, have been interpreted as icons of a fertility deity whose cult was widespread in the Mediterranean area and in western Asia from Neolithic times onward. This hypothesis is strengthened by the fact that the goddess was apparently associated with the bull—a feature also found in the ancient religions farther west.

The Harappā culture

Religion in the Indus Valley civilization. The Harappā culture (often called the Indus Valley civilization), located in modern Pakistan, has produced much evidence of the cult of the goddess and the bull. Figurines of both occur, with the goddess being more common than the bull. The bull, however, appears more frequently on the many steatite seals. A horned deity, possibly with three faces, occurs on a few seals, and on one seal he is surrounded by animals. A few male figurines in hieratic (sacerdotal) poses and one apparently in a dancing posture may represent deities. No building has been discovered at any Harappan site that can be positively identified as a temple, but the Great Bath at Mohenjo-daro was almost certainly used for ritual purposes, as were the ghats (bathing steps on riverbanks) attached to later Hindu temples. The presence of bathrooms in most of the houses and the remarkable system of covered drains indicate a strong concern for cleanliness that may have been related to concepts of ritual purity as well as to ideas of hygiene.

Sacred trees

Many seals show religious and legendary themes that cannot be interpreted with certainty. There is clear evidence, however, of the worship of sacred trees or of the divinities believed to reside in them. The bull is often depicted standing before a sort of altar, and the horned god has been interpreted, perhaps overconfidently, as a prototype of the Hindu god Śiva. Small conical objects appear to be phallic emblems that are also connected with Śiva in later Hinduism, although they may have been pieces used in board games. Other interpretations of the remains of the Harappā culture are more speculative and, if accepted, would indicate that many features of later Hinduism were already in existence 4,000 years ago. The fact that Harappans buried their dead with grave deposits, a practice not followed by the later Hindus, suggests that they had some belief in an afterlife.

Survival of archaic religious practices. Some elements of the religious life of current and past folk religions—notably sacred animals, sacred trees, especially the pipal (*Ficus religiosa*), and the use of small figurines for cult purposes—are found in all parts of India and may have been borrowed from pre-Aryan civilizations. On the other hand, these figures are also commonly encountered outside of India, and therefore they may have originated independently in Hinduism as well.

THE VEDIC PERIOD (2ND MILLENNIUM–7TH CENTURY BC)

The Aryans of the early Vedic period left few material remains, but they left a very important literary record called the Rigveda. Its 1,028 hymns are distributed throughout 10 books, of which the first and the last are the most recent. A hymn usually consists of three sections: it begins with an exhortation that is followed in the main part by praise of the deity, prayers, and imploration, with frequent references to the deity's mythology, and finishes with a specific request.

The Rigveda ("Wisdom of the Verses") is not a unitary work, and its composition may have taken several centuries. In its form at the time of its final edition it reflects a well-developed religious system. The date commonly given for the final recension of the Rigveda is 1000 BC. During the next two or three centuries the Rigveda was supplemented by three other Vedas and, still later, by Vedic texts called the Brāhmaṇas and the Upanishads (see below *Sacred texts: Vedas*).

CHALLENGES TO BRAHMANISM (7TH–2ND CENTURY BC)

The century from about 550 BC onward was a period of great change in the religious life of India. This century saw the rise of breakaway sects of ascetics who denied the authority of the Vedas and of the Brahmans and who followed founders claiming to have discovered the secret of obtaining release from transmigration. By far the

C.M. Naiu



Aspects of a soma sacrifice performed in Pune (Poona), India, on behalf of a traditional Brahman, following the same ritual used in 500 BC, an unusual example of the continuance of the Vedic tradition. (Though Vedic rites are still occasionally performed, photographic documentation of a complete and ritually correct ceremony is rare.) (Top) Group of 16 priests (four for each of the four Vedas) swear to the common interest of the sacrificer before the beginning of the ritual, their hands outstretched toward the fire (*āhavanīya*). (Bottom) Priests bring in fire to light the sacrificial fire in the newly constructed altar, taken from the household fire of the sacrificer, shown with his wife in the middle of the group.

most important of these were Siddhārtha Gautama, called the Buddha, and Vardhamāna, called Mahāvīra ("Great Hero"), the great teacher of Jainism (see also BUDDHISM, THE BUDDHA AND; JAINISM). There were many other heterodox teachers who organized bands of ascetic followers, and each group followed a specific code of conduct. They gained considerable support from ruling families and merchants. The latter were growing in wealth and influence, and many of them were searching for alternative forms of religious activity that would give them a more significant role than did orthodox Brahmanism or that would be less expensive to support.

The scriptures of the new religious movements throw some light on the popular religious life of the period. The god Prajāpati was widely believed to be the highest god and the creator of the universe, with Indra, known chiefly as Śakra ("the Mighty One"), second to him in importance. The Brahmans were very influential, but opposition had developed to their large-scale animal sacrifices—on both philosophical and economic grounds—and their pretensions to superiority by virtue of their birth were questioned. The doctrine of transmigration was by then generally accepted, although a group of outright materialists denied the survival of the soul after death. The ancestor cult, part of the Indo-European heritage, was retained almost universally, at least by the higher castes. Popular religious life largely centred around the worship of local fertility divinities (yaksha), snake-spirits (naga), and other minor spirits in sacred places and groves (*caitya*). Although these sacred places were the main centres of popular religious life, there is no evidence of any buildings or images associated with them, and it appears that neither temples nor large icons existed at the time.

Around 500 BC asceticism became widespread, and increasing numbers of intelligent young men "gave up the world" to search for release from transmigration by achieving a state of psychic security. The orthodox Brahmanical teachers reacted to these tendencies by devising the doctrine of the four ashramas (*āśramas*, "abodes"), which divided the life of the twice-born after initiation into four stages: the brahmachari (celibate religious student); the gṛihastha (married householder); the vanaprastha (forest dweller); and the sannyasin (wandering ascetic). This attempt to keep asceticism in check and confine it to men of late middle age was never followed universally, but thereafter Hindu social theory centred on the concept of varnashramadharma, or the duties of the four classes (varna) and the four stages of life (ashramas), which formed the ideal that Hindus were encouraged to follow.

The doctrine of the four ashramas

The 3rd century BC was the period of the Mauryan empire, the first great empire of India. Its early rulers were heterodox, and Aśoka (reigned *c.* 265–238 BC), the third and most famous of the Mauryan rulers, was a professed Buddhist. Although there is no doubt that Aśoka's patronage of Buddhism did much to spread that religion, his inscriptions recognize the Brahmans as worthy of respect. Sentiments in favour of nonviolence (ahimsa) and vegetarianism, much encouraged by the heterodox sects, spread during the Mauryan period and were greatly encouraged by Aśoka. A Brahmanic revival appears to have occurred with the fall of the Mauryas. The orthodox religion itself was undergoing change at this time, however, because of the development of theistic tendencies that centred around the gods Vishnu and Śiva.

The beginning of devotional theism

Inscriptions, iconographic evidence, and literary references point to the emergence of devotional theism in the 2nd century BC. Several brief votive inscriptions refer to the god Vāsudeva, who by this time was widely worshiped in western India. At the end of the 2nd century, Heliodorus, a Greek ambassador from King Antialcidas of Taxila (in Pakistan), erected a large column in honour of Vāsudeva at Besnagar in Madhya Pradesh and recorded that he was a Bhāgavata, a term specifically used for the devotees of Vishnu. The identification of Vāsudeva with the old Vedic god Vishnu and, later, with Vishnu's incarnation, Krishna (Kṛṣṇa), was quickly accepted.

Near the end of the Mauryan period the first surviving stone images of Hinduism appear. Several large, simply carved figures survive, not representing any of the great gods but rather yakshas, or local chthonic divinities connected with water, fertility, and magic. The original locations of these images are uncertain, but they were probably erected in the open air in sacred enclosures. Temples are not clearly attested in this period by either archaeology or literature. A few fragmentary images thought to be those of Vāsudeva and Śiva, the latter in anthropomorphic form and in the form of a lingam, or phallic emblem, are found on coins of the 2nd and 1st centuries BC.



Brahman priest reading the unbound folio manuscript of a sacred text, a *śrauta-sūtra*, at a Vedic *yajña* (sacrifice). The animal skin is the sacrificer's seat, the wooden cups are for drinking the soma.
C.M. Natu

EARLY HINDUISM (2ND CENTURY BC–4TH CENTURY AD)

The centuries immediately preceding and following the dawn of the Common Era saw the recension of the two great Sanskrit epics, the *Rāmāyaṇa* and the *Mahābhārata* (the latter incorporating into it the *Bhagavadgītā*). Although it was the worship of Vishnu, incarnate as Krishna in the *Mahābhārata* and as Rāma in the *Rāmāyaṇa*, that developed significantly during this period (see below *Sacred texts: Epics and Purāṇas*), the god Śiva is active in the *Mahābhārata*, and the cult of Śiva developed alongside the cult of Vishnu.

The rise of the major sects: Vaishnavism, Śaivism, and Śāktism. The Vedic god Rudra gained in importance from the end of the Rigvedic period. In the Śvetāśvatara Upanishad, Rudra is for the first time called Śiva and is described as the creator, preserver, and destroyer of the universe. His followers are called on to worship him with devotion (bhakti). The tendency for the laity to form themselves into religious guilds, or societies—evident in the case of the yaksha cults, Buddhism, and Jainism—promoted the growth of devotional Vaishnavism and Śaivism. These local associations of worshipers appear to have been a principal factor in the spread of the new cults. Theistic ascetics are less in evidence at this time; but a community of Śaivite monks, the Pāśupatas, was also in existence by the 2nd or 3rd century AD.

The period between the fall of the Mauryan empire (*c.* 185 BC) and the rise of the Gupta (*c.* AD 320) was one of great change, with most of the area of Pakistan and parts of western India being conquered by a succession of invaders. India was opened to influence from the West as never before, not only by its invaders but by way of the sea through the flourishing trade with the Roman Empire. The effects of the new contacts were most obvious in art and architecture. The oldest freestanding stone temple in the subcontinent has been excavated at Taxila, near Rāwalpindi, Pak. During the 1st century BC the Gandhāra school of sculpture arose in the same region and made use of Hellenistic and Roman prototypes, mainly in the service of Buddhism. At that time Hindu temples proba-

Western influences

bly were made of wood because no remains of them are extant; however, literary evidence shows that they must have existed.

By the time of the early Gupta empire the new theism had been harmonized with the old Vedic religion, and two of the main branches of Hinduism were fully recognized. The Vaishnavas had the support of the Gupta emperors, who took the title *paramabhāgavata* ("supreme devotee of Vishnu"). Vishnu temples were numerous and the doctrine of Vishnu's avatars (incarnations) was widely accepted. Of the 10 incarnations of later Vaishnavism, however, only two seem to have been much worshiped in the Gupta period. These were Krishna, the hero of the *Mahābhārata,* who also begins to appear in his pastoral aspect as the cowherd and flute player, and the divine boar (Varāha), of whom several impressive images survive from the Gupta period.

The Śaivites were also a growing force in the religious life of India. The sect of Pāśupata ascetics, founded by Lakulīśa (or Nahulīśa), who lived in the 2nd century AD, is attested by inscriptions from the 5th century and is among the earliest of the sectarian religious orders of Hinduism. Representations of the son of Śiva, Skanda (also called Kārttikeya, the war god), appeared on Kushān coins as early as AD 100. Śiva's other son, the elephant-headed Gaṇeśa, patron deity of commercial and literary enterprises, did not appear until the 5th century. Very important in this period was Sūrya, the sun god, who had temples built in his honour, although in modern times he is little regarded by most Hindus. The solar cult had Vedic roots but later may have expanded under Iranian influence.

Rising importance of goddesses

Several goddesses began to gain importance in this period. Although goddesses had always been worshiped in local and popular cults, they play comparatively minor roles in Vedic religion. Lakṣmī or Śrī, goddess of fortune and consort of Vishnu, was worshiped before the beginning of the Christian era, and several lesser goddesses are attested from the Gupta period. But the cult of Durgā, the consort of Śiva, was only beginning to gain importance in the 4th century, and the large-scale development of Śāktism (devotion to the active, creative principle personified as the Mother Goddess) did not take place until medieval times.

The development of temples. The Gupta period (4th–6th century) saw the rapid development of temple architecture. Earlier temples were made of wood, but free-standing stone and brick temples soon appeared in many parts of India. By the 7th century, stone temples, some of considerable dimensions, were found in the Aryanized parts of the country. Originally, the design of the Hindu temple may have borrowed from the Buddhist precedent, for in some of the oldest temples the image was placed in the centre of the shrine, which was surrounded by an ambulatory path resembling the path around the Buddhist stupa (a religious building containing a relic). Nearly all surviving Gupta temples are comparatively small; they consist of a small cella (central chamber), constructed of thick and solid masonry, with a veranda either at the entrance or on all sides of the building. The earliest Gupta temples, such as the Buddhist temples at Sānchi, have flat roofs; however, the *sikhara* (spire), typical of the north Indian temple, was developed in this period and with time steadily was made taller. The massive and tall tower of the Buddhist temple of Buddh Gayā, which was in existence in the 7th century, represents the culmination of Gupta temple architecture.

The Buddhists and Jains had made use of artificial caves for religious purposes, and these were adapted by the Hindus. Hindu cave-temples, however, are comparatively rare, and none has been discovered from earlier than the Gupta period. In the Pallava site of Mahābalipuram, south of Madras, a number of small temples were carved in the 7th century from outcroppings of rock and represent some of the oldest religious buildings in the Tamil country.

The spread of Hinduism in Southeast Asia and the Pacific. Hinduism and Buddhism had an immense impact on the civilizations of Southeast Asia and contributed greatly to the development of a written tradition in that area. Around the beginning of the Christian era, Indian merchants in comparatively large numbers settled there, bringing Brahmans and Buddhist monks with them. These religious men were patronized by local chiefs, who converted to the new religion. The earliest material evidence of Hinduism in Southeast Asia comes from Borneo, where late 4th-century Sanskrit inscriptions testify to the performance of Vedic sacrifices by Brahmans at the behest of local chiefs. Chinese chronicles attest an Indianized kingdom in Vietnam two centuries earlier. The dominant form of Hinduism exported to Southeast Asia was Śaivism, though some Vaishnavism was also known there. Later, from the 9th century onward, Tantrism, both Hindu and Buddhist, spread throughout the region.

The civilizations of Southeast Asia developed forms of Hinduism and Buddhism that had distinctive local features and were attuned to the local cultures, but the framework of their religious life was essentially Indian. Stories from the *Rāmāyaṇa* and the *Mahābhārata* became widely known in Southeast Asia and are still popular there in local versions. The people of Bali (in Indonesia) still follow a form of Hinduism adapted to their own genius. Versions of the *Manu-smṛti* ("Laws of Manu") were taken to Southeast Asia and were translated and adapted to indigenous cultures until they lost most of their original content.

Claims of early Hindu contacts farther east are more doubtful. There is little evidence of the influence of Hinduism on China and Japan, except through Buddhism.

Indian religious influence in the Mediterranean world. Nearly as dubious as the question of Hindu influence on the religious life of the Far East is its influence on that of the ancient Mediterranean world. The Greek philosopher Pythagoras may have obtained his doctrine of metempsychosis (transmigration, or passage of the soul from one body to another) from India, mediated by Achaemenian (6th–4th century BC) Persia, but similar ideas were known in Egypt and were certainly present in Greece before the time of Pythagoras. The Pythagorean doctrine of a cyclic universe may also be derived from India, but the Indian theory of cosmic cycles is not attested in the 6th century BC. Nevertheless, it is known that Hindu ascetics occasionally visited Greece. The most striking similarity of Greek and Indian thought is the resemblance between the system of mystical gnosis (esoteric knowledge) described in the *Enneads* of the Neoplatonic philosopher Plotinus (3rd century AD) and that of the *Yoga-sūtras* attributed to Patañjali, an Indian religious teacher sometimes dated in the 2nd century AD. The Patañjali text is the older, and influence must be suspected, though the problem of mediation remains difficult because Plotinus gives no direct evidence of having known anything about Indian mysticism. Several Greek (*e.g.,* Clement of Alexandria) and Latin writers show considerable knowledge of the externals of Indian religions, but none gives any intimation of understanding their more recondite aspects.

Certain Vaishnava legends, especially those referring to the infant Krishna, bear some resemblance to those of Christianity, and claims have been made by both Hinduism and Christianity that the one influenced the other. There is, however, no definitive evidence for the priority of either one.

THE RISE OF DEVOTIONAL HINDUISM (4TH–11TH CENTURY)

The medieval period saw the growth of new devotional religious movements centred on hymnodists who taught in the popular languages of the time. The new movements probably began with the appearance of hymns in Tamil associated with two groups of poets, the Nāyanārs, worshipers of Śiva, and the Āḻvārs, devoted to Vishnu. The oldest of these date from the early 7th century, although passages of devotional character can be found in earlier strata of Tamil literature.

Tamil influences

The term bhakti, in the sense of devotion to a personal god, appears in the *Bhagavadgītā* and the Śvetāśvatara Upanishad. In these early sources it represents a devotion still somewhat restrained and unemotional. The new form of bhakti, associated with singing in the languages of the common people, was highly charged with emotion, and the relation of worshiper and divinity was often de-

scribed by the analogy of that of lover and beloved. This devotional poetry is characterized by a mystical fervour not found in the Upanishads and the *Bhagavadgītā*, in both of which, even when the object of meditation is conceived as a personal God, there is little expression of passion. The Tamil "saints," however, felt an intense love (Tamil: *anbu*) of a personal kind toward their god. They experienced overwhelming joy in his presence and deep sorrow when he did not reveal himself. Some of them felt a profound sense of guilt or inadequacy in the face of the divine. But the dominant emotion in these poems is one of joy, often expressing itself in song and dance. The poems have a strong ethical content and encourage the virtues of love, humility, and brotherhood. The ideas of these poets, spreading northward, probably were the origin of the growth of bhakti in northern India.

Antagonism between devotional cults and Buddhism

The devotional cults further weakened Buddhism, which had long been on the decline. From time to time Hindus, especially Śaivites, took aggressive action against Buddhism. At least two Śaivite kings—the Hephthalite invader Mihirakula (early 6th century) and the Bengal king Śaśāṅka (early 7th century)—are reported to have been active persecutors, destroying monasteries and killing monks. The philosophers Kumārila and Śaṅkara were also strongly opposed to Buddhism. In their journeys throughout India, their biographies claim, they vehemently debated with Buddhists and tried to persuade kings and other influential people to withdraw their support from Buddhist monasteries. Only in Bihār and Bengal, because of the patronage of the Pāla dynasty and some lesser kings and chiefs, did Buddhist monasteries continue to flourish. Buddhism in eastern India, however, was well on the way to being reabsorbed into Hinduism when the Muslims invaded the Ganges (Gaṅgā) Valley in the 12th century. The great Buddhist shrine of Buddh Gayā, the site of the Buddha's enlightenment, became a Hindu temple and remained as such until recent times.

At the end of its existence in India, Buddhism developed in a way that had some effect on Hinduism. Among the Buddhist Tantrists appeared a new school of preachers, often known as siddhas (those who have achieved), who sang their verses in the contemporary languages, early Maithili and Bengali. They taught that giving up the world was not necessary for release from transmigration and that by living a life of simplicity in one's own home one could achieve the highest state. This system, known as Sahajayāna ("the Vehicle of the Natural," or "the Easy Vehicle"), influenced both Bengali devotional Vaishnavism, which produced sects called Sahājiyā with similar doctrines, and the Nātha yogis (mentioned below), whose teachings influenced Kabir and other later bhakti teachers.

HINDUISM UNDER ISLĀM (11TH–19TH CENTURY)

The challenge of Islām and popular religion. The phase of Indian history marked by the domination of the Muslims in most of northern India saw great changes in Indian religion. The advent of Islām in the Ganges Basin at the end of the 12th century resulted in the withdrawal of royal patronage from Hinduism in much of the area. The attitude of the Muslim rulers toward Hinduism varied. Some, like Fīrūz Tughluq (ruled 1351–88) and Aurangzeb (ruled 1658–1707), were strongly anti-Hindu and enforced payment of jizya, a poll tax on unbelievers. Others, like the Bengali sultan Ḥusayn Shāh 'Alā' ad-Dīn (reigned 1493–1519) and the great Akbar (reigned 1556–1605), were well-disposed toward their Hindu subjects. Many temples, however, were destroyed by the more fanatical rulers. Conversion to Islām was more common in areas where Buddhism had once been strongest—modern Pakistan, Bangladesh, and Kashmir.

Hinduism on the eve of the Muslim occupation

On the eve of the Muslim occupation, Hinduism was by no means sterile in northern India, but its vitality was centred in the southern, Dravidian-speaking areas. Throughout the centuries, the system of class and caste had become more rigid; in each region there was a complex hierarchy of castes strictly forbidden to intermarry and interdine, controlled and regulated by secular powers who acted on the advice of the court Brahmans. The large-scale Vedic sacrifices had practically vanished, but simple domestic Vedic sacrifices continued, and new forms of animal, and sometimes vegetarian, sacrifice had appeared, especially connected with the cult of the Mother Goddess.

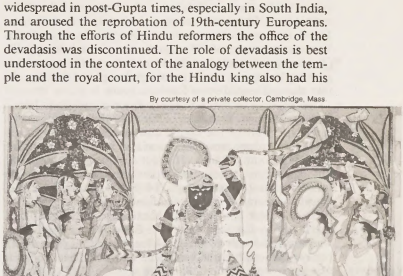
By that time, the main divinities of later Hinduism were worshiped. Rāma, the hero of the epic poem, had become the eighth avatar of Vishnu, and his cult was growing, although it was not yet as prominent as it later became. Similarly, Rāma's monkey helper, Hanumān, now one of the most popular divinities of India and the most ready helper in time of need, was rising in importance. Krishna was worshiped with his adulterous consort, Rādhā. Strange syncretic gods had appeared, such as Harihara, a combination of Vishnu and Śiva, and Ardhanārīśvara, a synthesis of Śiva and his shakti Pārvatī or Durgā.

Temple complexes. From the Gupta period onward Hindu temples tended to become larger and more prominent, and their architecture developed in distinctive regional styles. In northern India the best remaining Hindu temples are found in the Orissa region and in the town of Khajurāho in northern Madhya Pradesh. The best example of Orissan temple architecture is the Liṅgarāja temple of Bhubaneswar, built about 1000. The largest temple of the region, however, is the famous Black Pagoda, the Sun Temple (Sūrya Deuḷa) of Konārak, built in the mid-13th century. Its tower has long since collapsed, and only the assembly hall remains. The most important Khajurāho temples were built during the 11th century. Individual architectural styles also arose in Gujarāt and Rājasthān, but their surviving products are less impressive than those of Orissa and Khajurāho. By the end of the 1st millennium AD the South Indian style had reached its apogee in the great Rājarājeśvara temple of Thanjāvūr (Tanjore).

Hindu temples

In the temple the god was worshiped by the rites of puja (reverencing a sacred being or object) as though the worshipers were serving a great king. In the important temples a large staff of trained officiants waited on the god. He was awakened in the morning along with his goddess, washed, clothed and fed, placed in his shrine to give audience to his subjects, praised and entertained throughout the day, ceremoniously fed, undressed, and put to bed at night. Worshipers sang, burned lamps, waved lights before the divine image, and performed other acts of homage. The god's dancing girls (devadasis) performed before him at regular intervals, watched by the officiants and lay worshipers, who were his courtiers. These women, either the daughters of devadasis or girls dedicated in childhood, may have also served as prostitutes. The association of dedicated prostitutes with certain Hindu shrines can be traced back to before the Christian era. It became more widespread in post-Gupta times, especially in South India, and aroused the reprobation of 19th-century Europeans. Through the efforts of Hindu reformers the office of the devadasis was discontinued. The role of devadasis is best understood in the context of the analogy between the temple and the royal court, for the Hindu king also had his

By courtesy of a private collector, Cambridge, Mass.



"Four Priests Doing Puja Before the Image of Śrī Nāthajī," Rajasthani miniature painting from Nāthdwāra, *c.* 1830. In the Fogg Art Museum, Cambridge, Mass.

dancing girls, who bestowed their favours on his courtiers.

Parallels between the temple and the royal palace also were in evidence in the *rathayātras* (shrine processions). As on festival days, when the king issued from his palace and paraded around his city, escorted by courtiers, troops, and musicians, so also the god paraded around his city in a splendid procession, together with the lesser gods of the minor shrines. The god rode on a tremendous and ornate moving shrine (*ratha*), which was often pulled by large bands of devotees. *Rathayātras* still take place in many cities of India. The best-known is the annual procession of Jagannātha ("Juggernaut"), a form of Vishnu, at Puri, Orissa.

The wealth of the great temples

The great temples were (and still are) wealthy institutions. They were supported by the transfer of the taxes levied by kings on specific areas of the nearby countryside, by donations of the pious, and by the fees of worshipers. Their immense wealth was one of the factors that encouraged the Ghaznavid and Ghūrid Turks to invade India after the 11th century. They were controlled by self-perpetuating committees—whose membership was usually a hereditary privilege—and by a large staff of priests and temple servants under a high priest who wielded tremendous power and influence. The great walled temple complexes of South India were (and still are) small cities, containing the central and numerous lesser shrines, bathing tanks, administrative offices, homes of the temple employees, workshops, bazaars, and public buildings of many kinds. Directly and indirectly they played an important part in the economy, as they were among the largest employers and greatest landowners in their areas. They also performed valuable social functions because they served as schools, dispensaries, poorhouses, banks, and concert halls.

The Muslim occupation brought India into close contact with a different, more aggressive, religion. In such circumstances, the absence of a central religious authority in Hinduism was a source of strength. The *purohitas*, or family priests who performed the domestic rituals and personal sacraments for the lay people, continued to function, as did the thousands of ascetics. In Muslim-occupied territory the temples suffered the most. In the sacred cities of Vārānasi (Benares) and Mathurā, no large temple remains from any period before the 17th century. The same is true of most of the main religious centres of northern India, but not of the regions where the Muslim hold was less firm, such as Orissa, Rājasthān, and South India.

Sectarian movements. Before the time the Muslims invaded the subcontinent, the new forms of South Indian bhakti were spreading beyond the bounds of the Dravidian south. Certain Vaishnava theologians of the Pāñcarātra and Bhāgavata schools, including Rāmānuja, a Tamil Brahman who was for a time chief priest of the Vaishnava temple of Srirangam, near Tiruchchirāppalli (Trichinopoly), taught in the 11th century. They gave the growing Vaishnava bhakti cults a philosophical framework that also influenced some Śaivite schools.

Two other Vaishnava teachers deserve mention. Nimbārka, a Telugu Brahman of the 12th or 13th century, spread the cult of the divine cowherd and his favourite *gopī* (cowherdess, especially associated with the legends of Krishna's youth), Rādhā. His sect survives near Mathurā but has made little impact elsewhere. More important was Vallabha (Vallabhācārya; 1479–1531), who took the Vaishnava doctrine of grace and emphasized its erotic imagery. His sect is noteworthy because it stresses absolute obedience to the guru (teacher). Early in its existence it was organized with a hierarchy of senior monks (*gosvāmī*), many of whom became very rich. The Vallabhācārya sect

Vaishnava and Śaiva teachers

was once very influential in the western half of North India, but it declined in the 19th century, in part because of a number of lawsuits against the chief guru, the descendant of Vallabha.

The Śaiva sects also developed from the 10th century onward. In South India there emerged the school of Śaivasiddhānta, still one of the most significant religious forces in that region, and one that, unlike the school of Śaṅkara, does not admit the full identity of the soul and God. A completely monistic school of Śaivism appeared in Kashmir in the early 9th century. Its doctrines differ from those of Śaṅkara chiefly because it attributes personality to the absolute spirit, who is the god Śiva and not the impersonal *brahman*.

The unorthodoxy of the Liṅgāyats

An important and interesting sect, founded in the 12th century in the Kannada-speaking area of the Deccan, was that of the Liṅgāyats, or Virašaivas ("Heroes of the Śaiva Religion"). Its traditional founder, Basava, taught doctrines and practices of surprising unorthodoxy: he opposed all forms of image worship and accepted only the lingam of Śiva as a sacred symbol. Virašaivism rejected the Vedas, the Brahman priesthood, and all caste distinction. Several Liṅgāyat practices, now largely abandoned, such as the remarriage of widows and the burial of the dead, are deliberately antinomian.

An important development of Śaivism in North India was brought about by Gorakhnāth (Gorakṣanātha), who in the 13th century became leader of a sect of Śaivite ascetics known as Nātha ("Lord") from the title of their chief teachers. The Gorakhnāthis were particularly important as propagators of the practices of hatha-yoga, a form of yoga that requires complex and difficult physical exercises and that has become popular in the West. These yogis, who are still numerous, influenced the teaching of several of the bhakti poets.

Bhakti movements. The poets and "saints" of medieval bhakti appeared throughout India. Although all have their individual genius, the bhakti lyricists share a number of common features whatever their language. The Sanskrit education needed for authors of Sanskrit texts limited them largely to the Brahman class and thus put a definite stamp on them. Because bhakti poets could use any language, they might come from any class. They brought to their poetry a familiarity with folk religion unknown or ignored in the Sanskrit texts. The use of the spoken language, even though it was formalized, made possible the immediate expression of an unmediated vision that needed no further context; thus the lyrics are short, intensely personal, and precise. These works illustrate the localistic and reformist tendency evidenced throughout India in the vernacular literatures, especially in Tamil, Bengali, and Hindi. (See below *Vernacular literatures*.)

The origin of the new forms of Hinduism has been attributed to the influence of Islām, but the proposition that the rise of popular emotional bhakti was a response to Islām is impossible, for the practice of singing ecstatic hymns in the current local language was well-known in South India even before Muḥammad. All the features of this form of bhakti are found in the *Bhāgavata-Purāṇa* and in the commentaries of Rāmānuja. The earliest bhakti literature in a living Indo-Aryan language is from Mahārāshtra and was composed before Muslims occupied the area. Thus, passionate bhakti existed long before the Muslim conquest. However, the presence of rulers of alien faith and the withdrawal of royal patronage from the temples and Brahmanic colleges may have encouraged the spread of new, more popular forms of Hinduism. The psychological effect of the Muslim conquest may also have predisposed the people to accept the powerful teachings of the poets, but Islām was only a contributory factor in the spread of the new movements.

Much has been said about the synthesis of Hinduism and Islām in the period of Muslim dominance, but, as far as the Hindus were concerned, this was generally a matter of superficial observances. Thus, purdah (*pardā*), the strict seclusion of women, became commonplace among the Hindu upper classes of northern India, numerous Muslim social customs were adopted, and Persian and Arabic words entered the vocabularies of Indian languages. The fundamental theology of Hinduism, however, was unaffected by Islām, even in the teachings of such men as Basava and Kabīr, who may have been somewhat influenced by Muslim observances and social customs.

What synthesis did take place came from the Muslims, most of whom were Indian by blood. In Hindi, Bengali, Gujarati, Punjabi, and Marathi there is much poetic literature, written by Muslims and commencing with the Islāmic invocation of Allāh, which nevertheless betrays strong Hindu influence. Thus, there are texts that proclaim Krishna as being in the line of the prophets of

The Hindu influence on Islām

Islām and as the teacher of the unity of God. Much mystical poetry, though written by authors with Muslim names, uses Hindu imagery and Hindu terminology. This literature originated in the accommodating character of early Indian Ṣūfism, which, well before Kabīr, proclaimed that Muslim, Christian, Jew, Zoroastrian, and Hindu were all striving toward the same goal and that the outward observances that kept them apart were false. Some of the Indian Ṣūfīs were greatly influenced by Hindu customs. For example, a school of Kashmir Ṣūfīs, whose members call themselves rishis, after the legendary Hindu sages (*ṛṣi*), respect and repeat the verses of Lāl Ded and are strict vegetarians.

Syncretic tendencies were encouraged by tolerant Muslim rulers, and these tendencies reached their zenith in the reign of Akbar (1556–1605), who took a great interest in the religion of his Hindu subjects, favoured vegetarianism, and tried to establish a single, all-embracing religion for his empire. Although the efforts of Akbar failed, they influenced India for more than 50 years after his death. The orthodox Muslim theologians had long been complaining about the growth of heresy, however, and the emperor Aurangzeb (reigned 1658–1707) did all in his power to discourage it. Popular Muslim preachers throughout the 18th and 19th centuries worked to restore orthodoxy. Thus, syncretic tendencies virtually came to an end before the imposition of British power in the mid-18th century. Furthermore, British rule emphasized the distinctions between Hindu and Muslim and did not encourage efforts to harmonize the two religions.

THE MODERN PERIOD (19TH–20TH CENTURY)

Western Christian missionary activities

From their small coastal settlements in southern India, the Portuguese promoted Roman Catholic missionary activity and made converts, most of whom were of low caste; the majority of caste Hindus were unaffected. Small Protestant missions operated from the Danish factories of Tranquebar in Tamil Nādu and Serampore in Bengal, but they were even less influential. The British East India Company, conscious of the disadvantages of unnecessarily antagonizing its Indian subjects, excluded all Christian missionary activity from its territories. Indeed, the company continued the patronage accorded by indigenous rulers to many Hindu temples and forbade its Indian troops to embrace Christianity. The growing evangelical conscience in England brought this policy to an end with the renewal of the company's charter in 1813. The company's policy then became one of strict impartiality in matters of religion, and missionaries were allowed to work throughout its territory. Thus, Christian ideas began to spread.

Hindu reform movements. *Brahmo Samaj.* The pioneer of reform was Ram Mohun Roy. His intense belief in strict monotheism and in the evils of image worship began early and probably was derived from Islām, because at first he had no knowledge of Christianity. He later learned English and in 1814 settled in Calcutta, where he was prominent in the movement for encouraging education of a Western type. His final achievement was the foundation of the Brahmo Samaj ("Society of God") in 1828.

Roy outwardly remained a Hindu, wearing the sacred cord and keeping most of the customs of the orthodox Brahman; but his theology was surprisingly un-Indian. He was chiefly inspired by 18th-century Deism (rational belief in a transcendent creator god) and Unitarianism (belief in God's essential oneness), but some of his writing suggests that he was aware of the religious ideas of the Freemasons (a secret fraternity that espouses some Deistic concepts). Several of his friends were members of a Masonic lodge in Calcutta. His ideas of the afterlife are obscure, and it is possible that he did not believe in the doctrine of transmigration. Roy was one of the first higher-class Hindus to visit Europe, where he was much admired by the intelligentsia of Britain and France.

Debendranath Tagore and Keshab Chunder Sen

After Ram Mohun Roy's death, Debendranath Tagore became leader of the Brahmo Samaj, and under his guidance a more mystical note was sounded by the society. The third great leader of the Brahmo Samaj, Keshab Chunder Sen, was a vigorous reformer who completely abolished caste in the *samāj* and admitted women. As his theology became more syncretistic and eclectic, a schism developed, and the more conservative faction remained under the leadership of Tagore. Keshab's faction, the Brahmo Samaj of India, adopted as its scripture a selection of theistic texts gathered from all the main religions; at the same time, it became more Hindu in its worship, employing the *saṃkṛtrtana* (hymn-singing session) and *nagara-kīrtana* (street procession) of the Caitanya sect. In 1881 Keshab founded the Church of the New Dispensation (Naba Bidhan) for the purpose of establishing the truth of all the great religions in an institution that he believed would replace them all. When he died in 1884, the Brahmo Samaj began to decline, but it produced the greatest poet of modern India, Rabindranath Tagore (1861–1941), son of the second of its great leaders, Debendranath Tagore.

Rift between liberal and conservative factions

Arya Samaj. A reformer of different character was Dayanand Sarasvati, who was trained as a yogi but steadily lost faith in yoga and many other aspects of Hinduism. After traveling widely as an itinerant preacher, he founded the Arya Samaj in 1875, and it rapidly gained ground in the west of India. Dayanand rejected image worship, sacrifice, and polytheism and claimed to base his doctrines on the four Vedas as the eternal word of God. Later Hindu scriptures were judged critically, and many of them were believed to be completely evil. The Arya Samaj did much to encourage Hindu nationalism, but it did not disparage the knowledge of the West, and it established many schools and colleges. Among its members was the revolutionary Lala Lajpat Rai.

New religious movements. *Ramakrishna Mission.* The most important developments in Hinduism, however, did not arise primarily from the new *samāj*es. The mystic Ramakrishna, who was a devotee at a temple of Kālī called Dakṣiṇeśvar to the north of Calcutta, attracted a band of educated lay followers who spread his doctrines. As a result of his studies and visions, he came to the conclusion that "all religions are true" but that the religion of a person's own time and place was for him the best expression of the truth. Even idolatry met the needs of simple people and was not to be disparaged. Ramakrishna thus gave educated Hindus a basis on which they could justify the less rational aspects of their religion to a consciousness increasingly influenced by Western values.

Influence of the mystic Ramakrishna

Among the followers of Ramakrishna was Narendranath Datta, who became an ascetic after his master's death and assumed the religious name Vivekananda. In 1893 he attended the World's Parliament of Religions in Chicago, where his powerful personality and stirring oratory deeply impressed the gathering. After lecturing in the United States and England, he returned to India in 1897 with a small band of Western disciples. There he founded the Ramakrishna Mission, the most important modern organization of reformed Hinduism. Vivekananda, more than any earlier Hindu reformer, encouraged social service and the uplift of the downtrodden. Influenced by progressive Western political ideas, he set himself firmly against all forms of caste distinction and fostered a spirit of self-reliance in his followers. The Ramakrishna Mission has done much to spread a knowledge of its version of Hinduism outside India and now has branches in many parts of the world.

Theosophical Society. Another movement influenced in part by Hinduism is the Theosophical Society, which at one time exerted considerable influence. Founded in New York City in 1875 by Helena Blavatsky of Russia, it had as its original inspiration Kabbala (Jewish esoteric mysticism), Gnosticism (esoteric salvatory knowledge), and other forms of Western occultism. When Blavatsky went to India in 1879, her doctrines quickly took on an Indian character, and from her headquarters at Adyar she and her followers established branches in many cities of India.

The society survived serious accusations of charlatanry leveled against its founder and certain other leaders, and it reached the peak of its influence under its next important leader, Annie Besant, a reform-minded Englishwoman. Under her guidance, many Theosophical lodges were founded in Europe and the United States, and these helped to acquaint the West with the principles of Hinduism, if in a rather idiosyncratic form.

Aurobindo Ashram. Another modern teacher whose doctrines have had some influence outside India was Śrī Aurobindo, who began his career as a revolutionary. He later withdrew from politics and settled in Pondicherry, then a French possession. There he established an ashram, or *āśrama* (a retreat), and achieved a high reputation as a sage. His followers saw him as the first incarnate manifestation of the superbeings whose evolution he prophesied, and apparently he did not discourage this belief. After his death, the leadership of the Aurobindo Ashram was assumed by Mira Richard, a Frenchwoman who had been one of his disciples.

Other reform movements. Numerous other teachers have affected the religious life of modern India. Among them was the great Bengali poet Rabindranath Tagore, who was influenced by many currents of earlier religious thought, both Indian and other. Tagore was particularly popular in Europe and America around the time of World War I, and he did much to disseminate Hindu religious thought in the West.

Less important outside India, but much respected in India itself, especially in the Dravidian south, was Ramana Maharshi, a Tamil mystic who maintained almost complete silence. His powerful personality attracted a large band of devotees.

Swami Śivānanda, who had been a physician, established an ashram and an organization called the Divine Life Society near the sacred site of Rishikesh in the Himalayas. This organization has numerous branches in India and some elsewhere. His movement teaches more or less orthodox Vedānta, combined with both yoga and bhakti, but rejects caste and stresses social service.

Jiddu Krishnamurti represents the most attenuated form of Hinduism. He rejects all religious organization and makes no claim to special revelations or exceptional spiritual development; instead, he teaches self-realization through introspection and the abandonment of personal ambition.

Religion and the growth of nationalism

The struggle for independence. The Hindu revival and reform movements of the 19th and early 20th centuries were closely linked with the growth of Indian nationalism and the struggle for independence. The Arya Samaj strongly encouraged nationalism, and even though Swami Vivekananda and the Ramakrishna Mission were always uncompromisingly nonpolitical, their effect in promoting the movement for self-government is quite evident.

Religion and politics were joined in the career of Bal Gangadhar Tilak, an orthodox Mahārāshtrian Brahman who believed that the people of India could be aroused only by appeals couched in religious terms. Tilak used the annual festival of the god Gaṇeśa (Gaṇapati) for nationalist propaganda. His interpretation of the *Bhagavadgītā* as a call to action was also a reflection of his nationalism, and through his mediation the *Bhagavadgītā* became a stimulus to later leaders, including Mahatma Gandhi.

Hindu religious concepts also were enlisted in the nationalist cause in Bengal. In his historical novel *Ānandamaṭh*, the Bengali novelist Bankim Chandra Chatterjee described a band of martial ascetics at the time of the decline of the Mughal empire, who were pledged to free India from Muslim domination. These had as their anthem a stirring devotional song written in simple Sanskrit—"Bande Mātaram" ("I revere the Mother"). The Mother referred to is both the stern demon-destroying goddess Kālī and a personification of India. This song was soon adopted by the more extreme nationalists. Vivekananda emphasized the need to turn the emotion of bhakti toward the suffering poor of India. During his short career as a young revolutionary leader, Śrī Aurobindo made much use of "Bande Mātaram," and he called on his countrymen to strive for the freedom of India in a spirit of devotion. The bhakti of the medieval poets was thus enlisted in the cause of modern independence.

Mahatma Gandhi. Much influenced by the traditional bhakti of his native Gujarāt and fortified by Christian and other religious literature that encouraged similar attitudes, Mahatma Gandhi, the most important leader for independence, appeared to his followers as the quintessence of the Hindu tradition. His austere celibate life was one that the Indian laity had learned to respect implicitly. Gandhi's message reached a wider public than that of any of the earlier reformers.

The Western element in Gandhi's ideology has often been exaggerated. His doctrine of nonviolence can be found in many Hindu sources, although his beliefs were much strengthened by Christian ethical literature and especially by the later writings of Leo Tolstoy. His political technique of passive resistance, satyagraha, also has Indian precedents, although in this he was influenced by Western writers such as the American Henry David Thoreau. The chief innovations in Gandhi's philosophy were his belief in the dignity of manual labour and the equality of women. Precedents for both of these can be found in the writings of some 19th-century reformers, but they have little basis in earlier Indian thought. In many ways Gandhi was a traditionalist. His respect for the cow—which he and other educated Indians rationalized as the representative of Mother Earth—was a factor in the failure of his movement to attract large-scale Muslim support. His insistence on strict vegetarianism and celibacy among his disciples, in keeping with the traditions of Vaishnava ascetic ethics, also caused difficulty among some of his followers. Still, the success of Gandhi represented a political culmination of the movement of popular bhakti begun in South India early in the Christian era.

Gandhi as innovator and traditionalist

The mantle of Mahatma Gandhi fell on Vinoba Bhave, one of his most devoted Mahārāshtrian supporters. For some years after independence Vinoba led a campaign of social service that culminated in the *bhūdān* (land-giving) movement, which persuaded many landowners and wealthy peasants to give fields to landless labourers. This movement had some small success in rural areas, but it gradually lost momentum. Although the memory of Gandhi continues to be revered by most Indians, his policies and principles carry little weight. The great bulk of social service is performed by government agencies rather than by voluntary bodies, whether Gandhian or other.

The religious situation after independence. The increase of nationalism, after the division of India into India and Pakistan in 1947, led to a widening of the gulf between Hindus and Muslims. In the early 1970s it was fashionable in Indian circles to paint the relations of the two religions in earlier centuries as friendly, blaming alien rule for the division of India. In Pakistan the tendency has been to insist that Hindus and Muslims have always been "two nations," even though the Hindus were happy under their Muslim rulers. Neither position is entirely correct. In earlier times there was much mutual influence. But the conservative and rigid moralistic element in Indian Islām gained the upper hand long before British power was consolidated in India, and Islāmic influence on Hinduism remained superficial.

Among the pioneers of nationalism, Tilak glorified the Mahārāshtrian hero Śivājī as the liberator of his country from the alien yoke of the Mughals; and Bankim Chandra Chatterjee's militant ascetics, who pledged to conquer and expel the Muslims, sang a battle hymn that no orthodox Muslim could repeat. British rulers of India did little or nothing to lessen Hindu–Muslim tension, and their policy of separate electorates for the two communities worsened the situation. Many leaders of the Indian National Congress movement, such as Jawaharlal Nehru, carried their Hinduism lightly and favoured a secular approach to politics. The majority, however, followed the lead of Gandhi, whose insistence on Hindu values discouraged Muslims from joining his movement, despite the fact that at his prayer meetings he recited passages from the Qur'ān as well as from Hindu and Christian scriptures. To the right of the Congress politically, the Hindu Mahasabha was equally nationalistic, but its explicitly Hindu nationalism was not opposed to nonviolence in its drive to establish a Hindu state in India.

The transfer of power in 1947 was accompanied by slaughter and pillage of huge proportions. Millions of Hindus left their homes in Pakistan for India, and millions of Muslims migrated in the opposite direction. The tension culminated in the assassination of Gandhi by a Hindu fanatic in January 1948.

India as a secular state

The policy of the Indian government was to establish a secular state, and the successive Congress governments have broadly kept to this policy. The governments of the Indian states, however, have not been so restricted by constitutional niceties. Some state governments have introduced legislation of a specifically Hindu character.

On the other hand, the Congress government has passed legislation more offensive to Hindu traditional prejudices than anything that any British Indian government would have dared to enact. All forms of discrimination against "untouchables" (now usually referred to by euphemisms such as "harijans," or "people of God," instead of the British euphemism "scheduled castes") are forbidden, although it has been impossible to enforce the law in every case. A great blow to conservatism was dealt by legislation in 1955 and 1956 that gave full rights of inheritance to widows and daughters, enforced monogamy, and permitted divorce on quite easy terms. The 1961 law forbidding dowries further undermined traditional Hinduism. Although the dowry has long been a tremendous burden to the parents of daughters, the strength of social custom is such that the law cannot be fully enforced.

The social structure of traditional Hinduism is slowly crumbling in the cities. Intercaste and interreligious marriages are becoming more frequent among the educated, although some aspects of the caste system show remarkable vitality, especially in the matter of appointments and elections. The bonds of the tightly knit Hindu joint family are also weakening, a process helped by legislation and the emancipation of women. The professional priests, who perform rituals for lay people in homes or at temples and sacred sites, complain of the lack of custom, and their numbers are diminishing.

Nevertheless, Hinduism is far from dying. Mythological films, once the most popular form of entertainment, are enjoying a renaissance. Organizations such as the Ramakrishna Mission flourish and expand their activities. New teachers appear from time to time and attract considerable followings. Militant fundamentalist Hindu organizations such as the Society for the Self-Service of the Nation (Rashtriya Svayamsevak Sangh; RSS) are steadily growing. Such movements can be seen as the cause or the result, or both, of persistent outbreaks of communal religious violence involving Hindus and Sikhs in North India, Tamil Hindus and Sri Lankan Buddhists in Sri Lanka, Tamil extremists and moderates in Tamil Nādu, and, still everywhere, Hindus and Muslims.

The adaptability of Hinduism to changing conditions is illustrated by the appearance in the Hindu pantheon of a new divinity, of special utility in an acquisitive society. This is the goddess Santoṣi Mātā, first worshiped widely by women in many cities of Uttar Pradesh and now worshiped throughout India, largely as the result of a popular mythological film about her birth and the origin of her worship. The new goddess was unheard-of a few years ago and has no basis in any Purāṇic myth. Propitiated by comparatively simple and inexpensive rites performed in the home without the intervention of a priest, Santoṣi, it is believed, grants practical and obvious blessings, such as a promotion for a needy, overworked husband, a new radio, or even a refrigerator. News of Santoṣi's blessings is passed from housewife to housewife, and even moderately well-educated women have become her devotees.

The appearance of a new deity: Santoṣī Mātā

On both the intellectual and the popular level, Hinduism is thus in the process of adapting itself to new values and new conditions that have been brought about by mass education and industrialization and is responding to 20th-century challenges.

Hinduism outside India. Since the latter part of the 19th century large colonies of Hindu migrants have been established in East Africa, Malaysia, the islands of the Pacific and the Indian Ocean, and some of the islands of the West Indies. These migrants have taken their religion with them and have adhered to it faithfully for several generations. In recent years they have been aided by Hindu missionaries, chiefly from the Arya Samaj or the Ramakrishna Mission. Since World War II many Hindus have also settled in the United Kingdom. Most of these migrants, however, are comparatively uneducated, and their religion has made little impression on the people among whom they live. They also have made no serious attempts to gain converts. Yet, one of the most striking aspects of contemporary Western culture is its readiness to accept Eastern religious ideas in a way that is unprecedented since the days of the Roman Empire. A recent manifestation of the spread of Indian religious attitudes in the Western world is the Hare Krishna cult, officially known as the International Society for Krishna Consciousness, with its principal office in Los Angeles. This is essentially a bhakti movement, broadly following the precedents of Caitanya. Since its foundation by a Hindu sannyasi, A.C. Bhaktivedanta (Swami Prabhupāda), in 1966, its growth has been surprising, and *saṅkirtana* (devotional singing and dancing) can be seen in the streets of New York City and London, performed by young men and women from Christian or Jewish homes wearing dhotis and saris. These manifestations are part of a process that began in 1784 with the first English translation of a Hindu religious text, Charles Wilkins' version of the *Bhagavadgītā*.

Hinduism is not by nature a proselytizing religion, however, in part because of its inextricable roots in the social system and the land of India. In recent years, many new gurus, such as Bhagwan Shree Rajneesh and Satya Sai Baba, have been successful in making converts in Europe and the United States. The very success of these gurus, however, has produced material profits that many people regard as incompatible with the ascetic attitude appropriate to a Hindu spiritual leader; in some cases, the profits have led to notoriety and even legal prosecution. In addition, the self-proclaimed conversion to questionable forms of "Hinduism" by popular singers and film stars has tended both to increase the glamour and to diminish the respectability of these new forms of Orientalism. That Hinduism is flourishing in India is obvious; that it has made, and can continue to make, a genuine contribution to Western religious thought is undeniable; that the invasion of the gurus is a part of that contribution is highly debatable. (A.L.B./J.A.B.v.B./W.Do.)

Sacred texts

VEDAS

Importance of the Vedas. The Veda, meaning "Knowledge," is a collective term for the sacred scriptures of the Hindus. Since about the 5th century BC, the Veda has been considered to be the creation of neither human nor god; rather, it is regarded as the eternal Truth that was in ancient times directly revealed to or "heard" by gifted and inspired seers (rishis) who transcribed it into the most perfect human language, Sanskrit. Although most of the religion of the Vedic texts, which revolves around rituals of fire sacrifice, has been eclipsed by Hindu doctrines and practices, the absolute authority and sacredness of the Veda remains a central tenet of virtually all Hindu sects and traditions. Even today, as it has been for several millennia, parts of the Veda are memorized and recited as a religious act of great merit.

The components of the Veda. The Veda is the product of the Aryan invaders of the Indian subcontinent and their descendants, although the original inhabitants (disdainfully called *dāsyus*, or "slaves," in the Veda) may very well have exerted an influence on the final product. The Veda represents the particular interests of two classes of Aryan society, the priests (Brahmans) and the warrior-kings (Kṣatriyas), who together ruled over the far more numerous peasants (Vaiśyas).

The great works of Vedic literature

Vedic literature ranges from the Rigveda (Ṛgveda; *c.* 1400 BC) to the Upanishads (Upaniṣads; *c.* 1000–500 BC). This literature provides the sole documentation for all Indian religion before Buddhism and the early texts of classical Hinduism. Because it is the literature of a ruling class, it probably does not represent all the myths and cults of the early Indo-Aryans, let alone those of the non-Aryans.

The most important texts are the four collections (Saṃhitās) known as the Veda or Vedas (*i.e.*, "Book[s] of Knowledge"): the Rigveda ("Wisdom of the Verses"), the Yajurveda ("Wisdom of the Sacrificial Formulas"), the Sāmaveda ("Wisdom of the Chants"), and the Athar-

vaveda ("Wisdom of the Atharvan Priests"). Of these, the Rigveda is the oldest.

In the Vedic texts following these earliest compilations, the Brāhmaṇas (discussions of the ritual), Āraṇyakas (books studied in the forest), and Upanishads (secret teachings concerning cosmic equations), the interest in the early Rigvedic gods wanes, and they become little more than accessories to the Vedic rite. Polytheism begins to be replaced by a sacrificial pantheism of Prajāpati ("Lord of Creatures"), who is the All. In the Upanishads Prajāpati merges with the concept of *brahman*, the supreme reality and substance of the universe (not to be confused with the Hindu god Brahmā), replacing any specific personification, thus transforming the mythology into abstract philosophy.

Together, the components of each of the four Vedas—the Saṃhitās, Brāhmaṇas, Āraṇyakas, and Upanishads—constitute the revealed scripture of Hinduism, or the Sruti (*Śruti;* "Heard"). All other works—in which the actual doctrines and practices of Hindus are encoded—are recognized as having been composed by human authors and are thus classed as Smriti (*Smṛti,* "Remembered"). The categorization of Veda, however, is capable of elasticity. First, the Sruti is not exactly closed; Upanishads, for example, have been composed until recent times. Second, the texts categorized as Smriti inevitably claim to be in accord with the authoritative Sruti, and thus worthy of the same respect and sacredness. For Hindus, the Veda is a symbol of unchallenged authority and tradition.

The Rigveda. The religion reflected in the Rigveda is a polytheism mainly concerned with the propitiation of divinities associated with the sky and the atmosphere. Of these, the Indo-European sky father Dyaus was by then little regarded. More important were such gods as Indra, Varuṇa (guardian of the cosmic order), Agni (the sacrificial fire), and Sūrya (the Sun).

The soma sacrifice in the Rigveda

The main ritual activity referred to in the Rigveda is the soma sacrifice. Soma was a hallucinogenic beverage prepared from a now-unknown plant; recently it has been suggested that the plant was a mushroom and that later another plant was substituted for the agaric fungus, which had become difficult to obtain. The Rigveda contains a few clear references to animal sacrifice, which probably became more widespread later. There is some doubt whether the priests formed a separate class of society at the beginning of the Rigvedic period. If they did so, the prevailingly loose boundaries of class made it possible for a man of nonpriestly parentage to become a priest. By the end of the period, however, they had become a separate class of specialists, the Brahmans (Brāhmaṇas), who claimed superiority over all the other social classes, including the Rājanyas (later Kṣatriyas), the warrior-kings.

The Rigveda contains little about birth rituals, but the rites of marriage and disposal of the dead were basically the same as in later Hinduism. Marriage was an indissoluble bond cemented by a lengthy and solemn ritual centring on the domestic hearth. The funeral rites of the rich included cremation, although other funeral forms were also practiced. An interesting reference in one hymn shows that the wife of the dead man lay down beside him on the funeral pyre but was called upon to return to the land of the living before it was lighted. This may have been a survival from an earlier period when the wife was actually cremated with the husband, a custom that was revived in later times.

Among other features of Rigvedic religious life that were important for later generations were the munis. The muni was apparently a sort of shaman (a religious personage having healing and psychic transformation powers), trained in various magic arts and believed to be capable of supernatural feats, such as levitation. He was particularly associated with the god Rudra, a deity connected with mountains and storm and more feared than loved. Rudra developed into the Hindu god Śiva, and his prestige increased steadily. The same is true of Vishnu, a minor solar deity in the Rigveda, who later became one of the most important and popular divinities of Hinduism.

One of the favourite myths of the Aryans was one that attributed the origin of the cosmos to the god Indra, after he had slain the great dragon Vṛtra, a myth very similar to one known in early Mesopotamia. With time, such tales were replaced by more abstract theories that are reflected in several hymns of the late 10th book of the Rigveda. These speculative tendencies were the beginnings of the persistent effort of Indian philosophers to reduce all things to a single basic principle.

Elaborations of text and ritual: the later Vedas. The chronology of later Vedic developments is extremely vague, but it probably encompasses the period from 1000 to 500 BC, which are the dates of the Painted Grayware strata in the archaeological sites of the western Ganges Valley. These excavations reflect a culture still without writing but showing considerable advances in civilization. Nothing, however, has been discovered from sites of this period that throws much light on the religious situation, and historians still must rely on the following texts to describe this phase of the religion.

The other Saṃhitās

The Yajurveda and Sāmaveda. The Yajurveda and Sāmaveda are completely subordinate to the liturgy. The Yajurveda contains the lines, usually in brief prose, with which the executive priest (*adhvaryu*) accompanies his ritual manipulations, addressing the implements he handles and the offering he pours and admonishing other priests to do their invocations. The Sāmaveda is a collection of verses from the Rigveda (and a few new ones) that were chanted with certain fixed melodies.

The Atharvaveda. The Atharvaveda stands apart from other Vedic texts. It contains both hymns and prose passages and is divided into 20 books. Books 1–7 contain magical prayers for precise purposes: spells for a long life, cures, curses, love charms, prayers for prosperity, charms for kingship and Brahmanhood, and expiations for evil committed. They reflect the magical-religious concerns of everyday life and are on a different level than the Rigveda, which glorifies the great gods and their liturgy. Books 8–12 contain similar texts but also include cosmological hymns that continue those of the Rigveda and provide a transition to the more complex speculations of the Upanishads. Books 13–20 celebrate the cosmic principle (book 13) and present marriage prayers (book 14), funeral formulas (book 18), and other magical and ritual formulas. This text is an extremely important source of knowledge of practical religion and magic, particularly where it complements the one-sided picture of the Rigveda. Many rites are also laid down in the "Kauśikasūtra" (manual of the Kauśika family of priests) of the Atharvaveda.

The Brāhmaṇas and Āraṇyakas. Attached to each Saṃhitā was a collection of explanations of the rituals, called a Brāhmaṇa, which often relied on mythology to trace the origins and importance of individual ritual acts. Although they were not manuals or handbooks in the manner of the later Śrauta Sutras, the Brāhmaṇas do contain some detail about the performance and meaning of Vedic sacrificial rituals and are invaluable sources of information about the Vedic religion.

In these texts the sacrifice is the very centre of cosmic processes, all human concerns, and religious desires and goals. It is through the sacrifice that the cosmos continues in its cycles and that human beings obtain the goods of life and a birth in heaven in the next world. The ritual was thought to have such effects on the visible and invisible worlds because of homologies, or connections (*bandhus*), that were said to lie between the components and phases of the ritual and corresponding parts of the universe. The universalization of the dynamics of the ritual into the dynamics of the cosmos was depicted as the sacrifice of the primordial deity, Prajāpati ("Lord of Creatures"), who was perpetually regenerated by the sacrifice.

The lengthy series of rituals of the royal consecration, the *rājasūya,* emphasized royal power and endowed the king with a divine charisma, raising him, at least for the duration of the ceremony, to the status of a god. Typical of this period was the elaborate *aśvamedha,* the horse sacrifice, in which a consecrated horse was freed and allowed to wander at will for a year; it was always followed by the king's troops, who defended it from all attack until it was brought back to the royal capital and sacrificed in a very complicated ritual.

Vedic cosmic-sacrificial speculations continued in the

Āraṇyakas (forest books), which contain materials of two kinds: Brāhmaṇa-like discussions of rites not believed to be suitable for the village (hence the name "forest") and continuing visions of the relationship between sacrifice, universe, and man. The word *brahman*—the creative power of the ritual utterances, which is used to denote the creativeness of the sacrifice and which underlies ritual and therefore cosmic order—is prominent in these texts.

Vedic religion. *Cosmogony and cosmology.* In the Vedic literature there are different but not exclusive accounts of the origin of the universe. The simplest is that the creator built the universe with timber, as a carpenter builds a house. Hence there are many references to gods measuring the different worlds as parts of one edifice, atmosphere upon Earth, heaven upon atmosphere. Creation may be viewed as procreation: the personified Heaven, Dyaus (the word is related to the Greek Zeus), impregnates the Earth goddess, Pṛthivī, with rain, causing crops to grow on her. Quite another myth is recorded in the last (10th) book of the Rigveda: in the "Hymn of the Cosmic Man" ("Puruṣasūkta") it is said that the universe was created out of the parts of the body of a single cosmic man (Puruṣa) when his body was immolated and dismembered at the primordial sacrifice. There the four classes (varnas) of Indian society are referred to: the priest (Brahman) emerging from the mouth, the warrior (Rājanya) from the arms, the peasant (Vaiśya) from the thighs, and the servant (Śūdra) from the legs of the primeval victim. The "Puruṣasūkta" represents the beginning of a new phase, in which the sacrifice became more important and elaborate as cosmological and social philosophies were constructed around it.

Creation

In the same book of the Rigveda, mythology begins to be transformed into philosophy; for example, "in the beginning was the nonexistent, from which the existent arose" (Rigveda 10.72.2). Even the reality of the nonexistent is questioned: "then there was neither the nonexistent nor the existent" (Rigveda 10.129). Such cosmogonic speculations continue, particularly in the older Upanishads. Originally there was nothing at all, or Hunger, which then, to sate itself, creates the world as its food. Alternatively, the creator creates himself in the universe by an act of self-recognition, self-formulation, or self-formation. Or the one creator grows "as big as a man and a woman embracing" (Bṛhadāraṇyaka Upanishad I.4.3) and splits into man and woman, and in various transformations the couple create other creatures. In one of the last stages of this line of thought (Chāndogya Upanishad 6.2), the following account became fundamental to the ontology of the philosophical schools of Vedānta: in the beginning was the Existent, or *brahman,* which through heaven, Earth, and atmosphere (the triadic space) and the three seasons of summer, rains, and harvest (the triadic time) produced the entire universe.

The three worlds of heaven, atmosphere, and Earth

The Vedic texts generally regarded the universe as three layers of "worlds" (*loka*): heaven, atmosphere, and Earth. Heaven is that part of the universe where the sun shines and is correlated with sun, fire, and ether; the atmosphere is that part of the sky between heaven and Earth where the clouds insert themselves in the rainy season and is correlated with water and wind; Earth, a flat disk, like a wheel, is here below as the "holder of treasure" (*vasuṃdharā*) and giver of food. In addition to this tripartite pattern, there is also an ancient notion of duality, in which heaven is masculine and father and Earth is feminine and mother. Later texts present the conception that combinations and permutations of five elements (ether-space [*ākāśa*], wind [*vāyu*], fire [*agni*], water [*āpas*], and earth [*bhūmi*]) formed the universe.

Theology. Generally speaking, Vedic gods share many characteristics: several of them (Indra, Varuṇa, Vishnu) are said to have created the universe, set the Sun in the sky, and propped apart heaven and Earth. All of them are bright and shining, and all are susceptible to human praise. Some major gods were clearly personifications of natural phenomena, and for these deities no clearly delineated divine personalities were perceived.

Principal Vedic deities

The three most frequently invoked gods are Indra, Agni, and Soma. Indra, the foremost god of the Vedic pantheon, is a god of war and rain. Agni (a cognate of the Latin *ignis*) is the deified fire, particularly the fire of sacrifice, and Soma is the intoxicating or hallucinogenic drink of the sacrifice, or the plant from which it is pressed; neither is greatly personified.

The principal focus of Vedic literature is the sacrifice, which in its simplest form can be viewed as a ritualized banquet to which a god is invited to partake of a meal shared by the sacrificer and his priest. The invocations mention, often casually, the past exploits of the deity. The offered meal gives strength to the deity to repeat his feat and to aid the sacrificer.

The myth of Indra killing the dragon Vṛtra has many levels of meaning. Vṛtra prevents the monsoon rains from breaking. Because the monsoon is the greatest single factor in Indian agriculture, the event celebrated in this myth impinges on everyone's life. In the social circles represented in the Rigveda, however, the myth is cast in a warrior mold and the breaking of the monsoon is viewed as a cosmic battle. The entire monsoon complex is involved: Indra is the Lord of the Winds, the gales that accompany the monsoon; his weapons are lightning and thunderbolt, with which he lays Vṛtra low. To accomplish this feat he must be strengthened with soma. Simultaneously, he is the god of war and is invoked to defeat the non-Aryan *dāsyus,* the indigenous peoples referred to in the Vedas. These important concerns—the promptness and abundance of the rains, success in warfare, and the Aryan conquest of the land—all find their focus in Indra.

Because the Vedic gods were not fully anthropomorphic, their functions were subject to various applications and interpretations. Thus Indra, the greatest and most anthropomorphic god of the early Veda, was, in the view of the noble patrons of the Vedic poets, primarily a fighter and a warrior god invoked to bring booty and victory. Agriculturalists and hunters emphasized Indra's fecundity, celebrating his festivals to produce fertility, welfare, and happiness. Indra, however, was essentially a representative of useful force in nature and the cosmos and therefore was the great champion of an ordered and habitable world. His repeated victories over the snake-demon Vṛtra, the representative of obstruction and chaos, resulted in the separation of heaven and Earth (the support of the former and the stabilization of the latter), the rise of the Sun, and the release of the waters: in short, in the organization of the universe.

The notion of rita

Although morality is not an issue in Indra's myth, it is in those of the other principal Vedic deities. Central to ancient morality was the notion of rita (*ṛta*), the basic meaning of which appears to have been the truthfulness with which the alliance between humans (and between humans and gods) was observed—a quality necessary to maintain the physical and moral order of the universe. Varuṇa is an older sovereign god, who with Mitra (related to the Persian god Mithra) presides over the observance of the rita. Thus Varuṇa is a judge before whom a mortal may stand guilty, while Indra is a king who may support a mortal king. Typical requests that are made of Varuṇa are for forgiveness, for deliverance from evil committed by oneself or others, and for protection; Indra is prayed to for bounty, for aid against enemies, and for leadership against demons and *dāsyus.*

Distinct from both is Agni, the fire, who is observed in all his multifarious manifestations: in the sacrificial fire, in lightning, or hidden in the logs from which fire can be drilled. As the fire of sacrifice, he is the mouth of the gods and the carrier of the oblation, the mediator between the human and the divine orders. Agni is above all the good friend of the Aryans and is prayed to to strike down and to burn their enemies and to mediate between gods and men.

Among other Vedic gods, only a few stand out. One is Vishnu, important perhaps more in retrospect than in fact. He is famous for his "three strides," with which he traversed the universe, thus creating and possessing it. In his later mythology this pervasiveness, which invites identification with other gods, remains characteristic. His function as helper to the conqueror-god Indra is important.

Impersonality is increased by the prevalence of pairs and groups of gods. Thus Varuṇa and Mitra are members of the group of Ādityas (sons of Aditi, an old progenitrix),

who generally are celestial gods. They are also combined in the double god Mitra-Varuṇa. Indra and Vishnu are combined as Indra-Vishnu. There is also Rudra, an ambivalent god who is dreaded for his unpredictable attacks but is simultaneously benign insofar as he can restrain his attacks. Although there are many demons (rakshasas), no one god embodies the evil spirit; rather, many gods have their devil within, inspiring fear as well as trust. Among the perpetually beneficent gods are the Aśvins (horsemen), who are helpers and healers and often visit the needy. Almost otiose is the personified heaven, Dyaus, who most often appears literally as the sky, and often as day. As a person, he is coupled with Earth (in the god pair Dyāvā-Pṛthivī) as a father; Earth by herself is more predominantly known as Mother (Mātṛ). Apart from Earth, the other goddess of importance in the text of the Rigveda is Uṣas (Dawn), who brings in the day and thus is said to bring forth the Sun.

Vedic kathenotheism

In the later Vedic period the significance of the Rigvedic gods and their myths began to wane. The peculiar theism of the Rigveda, in which any one of several different gods might be hailed as supreme and attributes of one god could be transferred to another (called kathenotheism by the Vedic scholar F. Max Müller), stressed godhead more than individual gods. In the end this led to a pantheism of Prajāpati, the deified sacrifice or ritualized deity; with his consort Vāc (*i.e.,* the speech of ritual recitation), he is said to have begotten the world.

In the course of the Vedic period Puruṣa fused with the figure Nārāyaṇa ("Scion of Man") and with Prajāpati ("Lord of Creatures"), the patron of procreation in popular belief. In the speculative thought of the ritualists, Prajāpati came to the fore as the creator god and in many respects as the highest divinity, the immortal father even of the gods, whom he transcends, encompasses, and molds into one complex. As the One, the concentrated All, or Totality, Prajāpati was identified with the highest and most general categories. By a process of emanation and self-differentiation (by dividing himself), he created all beings and the universe. After this "creation," Prajāpati became the disintegrated and differentiated All of the phenomenal world and was exhausted. By means of a rite, he then reintegrated himself to prepare for a new phase of creativity. Because the purpose of the sacred act is the restitution of the organic structural norm, which ensures the ordered functioning of the universe, Prajāpati was identified with the rite. Thus, by identifying himself with Prajāpati, a sacrificer may temporarily reintegrate within himself what has been disintegrated, thereby restoring oneness and totality in himself and the universe.

Sin and evil

Ethical and social doctrines. In Vedic times, "sin" (*énas*) or evil (*pāpmán*) was put on a par with illness, enmity, distress, or malediction: it was conceived of as a sort of pollution that could be neutralized by ritual or devices for averting evil. A man might incur "sin" by any incorrect or improper behaviour, especially improper speech, and thus be guilty of *anṛta* (*i.e.,* any infidelity to fact or departure from what is true, real, and constitutes the established order) whether or not he had deliberately committed a crime. Other transgressions included making mistakes in sacrifices and coming into contact with corpses, ritually impure persons, or persons belonging to the lower classes of society. These acts were only rarely considered to be misdeeds against a god or violations of moral principles of divine origin, and the consciousness of guilt was much rarer than the fear of the evil consequences of sin, such as disease or untimely death. Sometimes, however, a god (Agni, the evil-devouring fire, or Varuṇa, the god of order, whose role included punishing and fettering the "sinner") was invoked to forgive the neglect or transgression or to release a man from their concrete results. More usually, however, these results were abrogated by means of purifications, such as the ceremonial use of water, and a variety of expiatory rites.

Importance of ritual purity

To the pure who earned ritual merits, the prospect of a safe "world" (*loka*) or condition was held out. The meticulous effort to purify oneself from every kind of evil also involved the observance of various customs regarding the avoidance of inauspicious occurrences—an endeavour called *śānti*. Ritual purity was the principal concern of the compilers of the manuals of dharma (religious law) that, belonging to the sacred tradition (Smriti; *i.e.,* remembered by human teachers), have contributed much to the special character of Hinduism. According to the authorities on dharma, ritual purity is: the first approach to dharma, the resting place of the Veda (*brahman*), the abode of prosperity (*śrī*), the favourite of the gods, and the means of clearing (soothing) the mind and of seeing (realizing) the atman in the body.

The sacred: nature, man, and God. The contact with the unseen or sacred included humankind's contributions by ritual acts to the maintenance of the universe—of which Vedic thinkers felt themselves an indissoluble part—and to the periodic regeneration, through sacrificial practices, of both the powers for good and the cosmic processes that make earthly life and welfare possible. The Vedic poets were deeply convinced that the world is an organized cosmos governed by order and truth (rita) and that it is always in danger of being damaged or destroyed by the powers of chaos (*asat*). This conviction found mythological expression in the continual conflict between gods (devas) and demoniac antigods (asuras).

Gods were conceived as presiding over certain provinces of the universe or as responsible for important cosmic or social phenomena. Their deeds are timeless and exemplary presentations of mythic events replete with power and universal, eternal significance. To reproduce themselves in time and thus retain their vitality and efficacy, mythical events need to be repeated—that is, celebrated and confirmed by means of the spoken word and ritual acts.

Vedic and Brahmanic rites. Vedic religion is primarily a liturgy differentiated in various types of ritual designed for almost any conceivable purpose. These rites are described in the texts in minute detail; theoretically, no operation, no gesture, no formula is meaningless or left to an officiant's discretion. On the basis of a complicated speculative system, all are explained and shown to be effective in the Brāhmaṇas. The often complicated ritual technique was devised mainly to safeguard human life and survival, to enable people to face the many risks and dangers of existence, to thwart the designs of human and superhuman enemies that cannot be counteracted by ordinary means, to control the unseen powers, and to establish and maintain beneficial relations with the supramundane sacred order. Belief in the efficacy of the rites is the natural consequence of the belief that all things and events are connected with or participate in one another. Hence it is also believed that a close correspondence exists between a sacred place—such as the sacrificial place of many Vedic rites, a place of pilgrimage, or a consecrated area (*maṇḍala,* "circle")—and a province of the universe or even the universe itself. These places represent, within the reach of the officiants, the universe or as much of it as is relevant. In such places, direct communication with other cosmic regions (heaven or underworld) is possible because they are said to be at the point of contact between this world and the "pillar of the universe," "the navel of the earth." The sacred place is (by virtue of a system of connections) identical with the universe in its various states of emanation from, reabsorption into, integration with, and disintegration from the sacred. This idea has as its corollary the possibility of ritually enacting the cosmic drama and, thus, of influencing, through the same system of connections, those events in the cosmos that continuously affect human weal and woe.

Cosmic role of sacred places

The Vedic ritual system is organized into three main forms. The simplest, and hierarchically inferior, type of Vedic ritualism is the *gṛhya,* or domestic ritual, in which the householder himself offers modest oblations into the one sacred household fire. The more ambitious, wealthy, and powerful married householder sets three or five fires and, with the help of professional officiants, engages in the more complex *śrauta* sacrifices. These require oblations of vegetable substances and, in some instances, of parts of ritually killed animals (mostly goats, but also sheep, cows, horses, and perhaps at one time human beings as well). Finally, at the highest level of Vedic ritualism are the sacrifices of soma, which can continue for days or even years

and whose intricacies and complexities are truly stunning.

In the major *śrauta* rites, requiring three fires and 16 priests or more, "the man who knows"—he who has an insight into the correspondences (*bāndhu*) between the mundane and cosmic phenomena and the eternal transcendent reality beyond them and who knows the meaning of the ritual words and acts—may set great cosmic processes in motion for the sake of human interests. In these rites, Brahman officiants repeat the mythic drama for the benefit of their patron, the "sacrificer," who temporarily becomes its centre and realizes through ritual symbolism his identity with the universe. Whatever magical elements may be involved in this ritual technique, its aim in establishing an efficacious contact with a transcendental order that is the source of all life and power is based on an essentially religious conception. Such officiants are firmly convinced of the efficacy of their rites: "the sun would not rise, were he [the officiant] not to make that offering; this is why he performs it" (Śatapatha Brāhmaṇa 2.3.1.5). The oblations should not be used to propitiate the gods or to thank them for favours bestowed, since the efficacy of the rites, some of which are still occasionally performed, does not depend on the will of the gods.

The Upanishads. With the last component of the Veda, the mystically oriented and originally esoteric texts known as the Upanishads, Vedic ritualism and the doctrine of the interconnectedness of separate phenomena was superseded by a new emphasis on knowledge alone—primarily knowledge of the ultimate identity of all phenomena, which merely appeared to be separate. The phase of Indian religious life roughly between 700 and 500 BC was the period of the beginnings of philosophy and mysticism marked by the Upanishads ("Sittings Near a Teacher"). Historically, the most important of these are the two oldest, the Bṛhadāraṇyaka ("Great Forest Text") and the Chāndogya (pertaining to the Chandogas, a class of priests who intone hymns at sacrifices), both of which are compilations that record the traditions of sages (rishis) of the period, notably Yājñavalkya, who was a pioneer of new religious ideas.

The primary motive of the Upanishads is a desire for mystical knowledge that would ensure freedom from "redeath." Throughout the later Vedic period, the idea that the world of heaven was not the end—and that even in heaven death was inevitable—had been growing. For Vedic thinkers, the fear of the impermanence of religious merit and its loss in the hereafter, as well as the fear-provoking anticipation of the transience of any form of existence after death, culminating in the much-feared repeated death (*punarmṛtyu*), assumed the character of an obsession. The means of escaping and conquering death and of attaining integral life devised in the Brāhmaṇas were of a ritual nature, but in one of the oldest Upanishads, the Bṛhadāraṇyaka (*c.* 10th–5th century BC), more emphasis was placed on the knowledge of the cosmic connection underlying ritual. When the doctrine of the identity of atman (the Self) and *brahman* was established in the Upanishads, the true knowledge of the Self and the realization of this identity was (by those sages who were inclined to meditative thought) substituted for the ritual method.

Development of atman-*brahman* doctrine from ritual speculation

In the following centuries, the main theories connected with the divine essence underlying the world were harmonized and synthetically combined, and the tendency was to extol one god as the supreme Lord and Originator (Īśvara), who is at the same time Puruṣa and Prajāpati and *brahman* and the inner Self (atman) of all beings. For those who worshiped him, he became the goal of identificatory meditation, which leads to complete cessation of phenomenal existence and becomes the refuge of those who seek eternal peace.

The period during which the Upanishads were composed was one of much social, political, and economic upheaval. Rural tribal society was disappearing, and the adjustments of the people to urban living under a monarchy probably provoked many psychological and religious responses. During this period many groups of mystics, world-renouncers, and forest-dwellers appeared in India, and these included the authors of the Upanishads. Among the more important practices and doctrines of these world-renouncers were asceticism and the concept of rebirth or transmigration.

The Rigveda shows few examples of asceticism, except among the munis (shamans). The Atharvaveda describes another class of religious adepts, or specialists, the *vrātyas*, particularly associated with the region of Magadha (west central Bihār). The *vrātya* was a wandering hierophant (one who manifested the Holy) who remained outside the regular system of Vedic religion. He traveled from place to place in a bullock cart with an apprentice and with a woman who appears to have been used for ritual prostitution. Flagellation and other forms of self-mortification seem to have been part of his routine. Efforts were made by the orthodox to bring the *vrātyas* into the Vedic system by special rituals of conversion, and it may be that these people helped to introduce non-Aryan beliefs and practices into Vedic religion. At the same time, the more complex sacrifices of the later Vedic period demanded purificatory rituals, such as fasting and vigil, as part of the preparations for the ceremony. Thus there was a growing tendency toward the mortification of the flesh.

The origin and the development of the belief in the transmigration of souls are very obscure. A few passages suggest that this doctrine was known even in the days of the Rigveda, but it was first clearly propounded in the earliest Upanishad—the Bṛhadāraṇyaka. There it is stated that normally the soul returns to Earth and is reborn in human or animal form. This doctrine of saṃsara (reincarnation) is attributed to the sage Uddālaka Āruṇi, who is said to have learned it from a Kṣatriya chief. In the same text, the doctrine of karma (actions), according to which the soul achieves a happy or unhappy rebirth according to its works in the previous life, also occurs for the first time, attributed to Yājñavalkya. Both doctrines appear to have been new and strange ones, circulating among small groups of ascetics who were disinclined to make them public, perhaps for fear of the orthodox priests. These doctrines must have spread rapidly, for in the later Upanishads and in the earliest Buddhist and Jain scriptures they are common knowledge.

SUTRAS, SHASTRAS, AND SMRITIS

The Vedāṅgas. Toward the end of the Vedic period, and more or less simultaneously with the production of the principal Upanishads, concise, technical, and usually aphoristic texts were composed about various subjects relating to the proper and timely performance of the Vedic sacrificial rituals. These were eventually labeled as Vedāṅgas ("Studies Accessory to the Veda").

Early scholarly disciplines

The intense preoccupation with the liturgy gave rise to scholarly disciplines that were part of the Vedic erudition. There were six such fields: (1) *śikṣā* (instruction), which explains the proper articulation and pronunciation of the Vedic texts. Different branches had different ways of pronouncing the texts, and these variations were recorded in *prātiśākhyas* (literally, "Instructions for the *śākhās*"—branches), four of which are extant; (2) *chandas* (metre), of which there remains only one late representative; (3) *vyākaraṇa* (analysis and derivation), in which the language is grammatically described—Pāṇini's famous grammar (*c.* 400 BC) and the *prātiśākhyas* are the oldest examples of this discipline; (4) *nirukta* (lexicon), which discusses and gives meanings for difficult words, represented by the *Nirukta* of Yāska (*c.* 600 BC); (5) *jyotiṣa* (luminaries), a system of astronomy and astrology used to determine the right times for rituals; and (6) *kalpa* (mode of performance), which studies the correct ways of performing the ritual.

Of special importance are the texts constituting the Kalpa Sutras (collections of aphorisms on the mode of ritual performance). The composition of these texts was begun around 600 BC by Brahmans belonging to the ritual schools (*śākhās*), each of which was attached to a particular recension of one of the four Vedas. A complete Kalpa Sutra contains four principal components: (1) a Śrauta Sutra, which establishes the rules for performing the more complex rituals of the Vedic repertoire; (2) a Śulba Sutra, which shows how to make the geometric calculations necessary for the proper construction of the ritual arena; (3) a Gṛhya Sutra, which explains the rules for performing the

domestic rites, including the life-cycle rituals (called the samskaras); and (4) a Dharma Sutra, which provides the rules for the conduct of life.

Society was ritually stratified in the four classes, each of which had its own dharma (law), or eternal norm of conduct. The ideal life was constructed through sacraments in the course of numerous ceremonies, performed by the upper classes, that carried the individual from conception to cremation in a series of complex rites. The Gṛhya Sutras show that in the popular religion of the time there were many minor divinities who are rarely mentioned in the literature of the large-scale sacrifices but who were probably far more influential on the lives of most people than were the great gods of Vedism.

Dharma Sutra and Dharma Shastra. Among the texts inspired by the Veda are the Dharma Sutras, or manuals on dharma, which contain rules of conduct and rites as they were practiced in a number of branches of the Vedic schools. Their principal contents address the duties of people at various stages of life or ashramas (studenthood, householdership, retirement, and asceticism); dietary regulations; offenses and expiations; and the rights and duties of kings. They also discuss purification rites, funerary ceremonies, forms of hospitality, and daily oblations. Finally, they even mention juridical matters. The more important of these texts are the sutras of Gautama, Baudhāyana, and Āpastamba. Although the direct relationship is not clear, the contents of these works were further elaborated in the more systematic Dharma Shastras, which in turn became the basis of Hindu law.

Hindu law

First among them stands the Dharma Shastra of Manu, also known as the *Manu-smṛti* ("Tradition of Manu"; *c.* AD 200), with 2,694 stanzas divided into 12 chapters. It deals with various topics such as cosmogony, definition of dharma, the sacraments, initiation and Vedic study, the eight forms of marriage, hospitality and funerary rites, dietary laws, pollution and purification, rules for women and wives, royal law, 18 categories of juridical matters, and finally more religious matters, including donations, rites of reparation, the doctrine of karma, the soul, and punishment in hell. Law in the juridical sense is thus completely embedded in religious law and practice. The framework is provided by the model of the four-class society. The influence of the Dharma Shastra of Manu has been enormous, as it provided Hindu society with its practical morality. For large parts of the Indian subcontinent, Manu's text—mediated by its commentaries, notably that of Medhātithi (9th century)—has been the law.

Second only to Manu is the Dharma Shastra of Yājñavalkya; its 1,013 stanzas are distributed under the three headings of good conduct, law, and expiation. Its commentary, *Mitākṣarā* of Vijñāneśvara (11th century), has extended its influence.

Smṛiti texts. The shastras are a part of the Smriti ("Remembered," or traditional) literature which, like the sutra literature that preceded it, stresses the religious merit of gifts to Brahmans. Because kings often transferred the revenues of villages or groups of villages to Brahmans, either singly or in corporate groups, the status and wealth of the priestly class rose steadily. In the *agrahāras,* as the settlements of Brahmans were called, they were encouraged to devote themselves to the study of the Vedas and the subsidiary studies associated with them; but many Brahmans also developed the sciences of the period, such as mathematics, astronomy, and medicine, while others cultivated literature.

Family law and the Smriti texts

The Smriti texts are binding to this day on orthodox Hindus, and until quite recently Hindu family law was based on them. Although there is evidence of divorce in early Indian history, by the Gupta period marriage was solemnized by lengthy sacred rites and was virtually indissoluble. Intercaste marriage was becoming rarer and more difficult, and child marriage and the rite of suttee were already in existence, although less frequent than they later became. One of the earliest definite records of a widow burning herself on her husband's pyre is found in an inscription from Eran, Madhya Pradesh, dated 510, but the custom had been followed sporadically long before this. From the 6th century AD onward, such occurrences became frequent in certain parts of India, particularly in Rājasthān.

EPICS AND PURĀṆAS

During the centuries immediately preceding and following the beginning of the Christian era, the recension of the two great Sanskrit epics, the *Mahābhārata* and the *Rāmāyaṇa,* took shape out of existing material such as heroic epic stories, mythology, philosophy, and above all the discussion of the problem of dharma. Much of the material of which the epics are composed dates far back into the Vedic period, while the rest continued to be added until well into the medieval period. It is conventional, however, to date the recension of the Sanskrit texts to the period from 300 BC to AD 300 for the *Mahābhārata* and to the period from 200 BC to AD 200 for the *Rāmāyaṇa.*

The Mahābhārata. The *Mahābhārata* ("Great Epic of the Bhārata Dynasty"), a text of some 100,000 verses attributed to the sage Vyāsa, was preserved both orally and in manuscript form for centuries. The central plot concerns a great battle between the five sons of Pāṇḍu, called the Pāṇḍavas (Arjuna, Yudhiṣṭhira, Bhīma, and the twins Nakula and Sahadeva), and the sons of Pāṇḍu's brother Dhṛtarāṣṭra. The battle eventually leads to the destruction of the entire race, save for one survivor who continues the dynasty. As each of the heroes is the son of a god (Indra, Dharma, Vāyu, and the Aśvins, respectively), the epic is deeply infused with religious implications. There are, moreover, many passages in which dharma is systematically treated, so that Hindus regard the *Mahābhārata* as one of the Dharma Shastras. Religious practice takes the form of Vedic ritual on official occasions, pilgrimage, and, to some extent, adoration of gods. Apart from the *Bhagavadgītā* (part of book 6) much of the didactic material is found in the Book of the Forest (book 3), in which sages teach the exiled heroes, and in the Book of Peace (book 12), in which the wise Bhīṣma expounds on religious and moral matters.

The Vedic gods have lost importance and survive as figures of folklore. Prajāpati of the Upanishads is popularly personified as the god Brahmā, who creates all classes of beings and dispenses boons. Of far greater importance is Krishna. In the epic he is a hero, a leader of his people, and an active helper of his friends. His biography as it is known later is not worked out; still, the text is the source of early Krishnaism. Not everywhere, and certainly not by everyone, is Krishna considered a god, and even as god his stature is superhuman rather than divine. He is occasionally, but not significantly, identified with Vishnu. Later, as one of the most important of the incarnations of Vishnu, Krishna undergoes a complex development as an incarnate god. In the *Mahābhārata* he is primarily a hero, a chieftain of a tribe, and an ally of the Pāṇḍavas, the heroes of the *Mahābhārata.* He accomplishes heroic feats with the Pāṇḍava prince Arjuna. Typically he helps the Pāṇḍava brothers to settle in their kingdom and, when the kingdom is taken from them, to regain it. In the process he emerges as a great teacher who reveals the *Bhagavadgītā,* the most important religious text of Hinduism. In the further development of the Krishna myth, this dharmic aspect recedes and makes way for an idyllic myth about Krishna's boyhood, when he plays with and loves young cowherd women (*gopīs*) in the village while hiding from an uncle who threatens to kill him. The influence of this theme on art has been profound. But there is a shadow side to this idyll. Even in the *Mahābhārata,* where it is often said that Krishna becomes incarnate in order to sustain dharma when it wanes and to combat adharma (forces contrary to dharma), he himself commits a number of deeds in direct violation of the warrior ethic and is indirectly responsible for the destruction of his entire family. This adharmic shadow is also cast in the Purāṇic idyll because the *gopīs* that he woos are the wives of other men.

Krishna in the *Mahābhārata*

Far remoter than the instantly accessible Krishna is Śiva, who also is hailed as the supreme god in several myths recounted of him, notably the Story of the Five Indras, Arjuna's battle with him, and his destruction of the sacrifice of Daksha. The epic is rich in information about sacred places, and it is clear that making pilgrimages and

bathing in sacred rivers constituted an important part of religious life. Occasionally these sacred places are associated with sanctuaries of gods. More frequent are accounts of mythical events concerning the particular place and enriching its sanctity. Numerous descriptions of pilgrimages (*tīrthayātrā*) give the authors opportunities to detail local myths and legends. In addition to these, countless edifying stories shed light on the religious and moral concerns of the age. Almost divine are the towering ascetics capable of fantastic feats, whose benevolence is sought and whose curses are feared.

The Rāmāyaṇa. The classical narrative of Rāma is recounted in the Sanskrit epic the *Rāmāyaṇa* by the sage Vālmiki, who is the traditional author of the epic. Rāma is deprived of the kingdom to which he is heir and is exiled to the forest with his wife Sītā and his brother Lakṣmaṇa. While there, Sītā is abducted by Rāvaṇa, the demon king of Laṅkā. In their search for Sītā, the brothers ally themselves with a monkey king whose general, Hanumān (who later became a monkey deity), finds Sītā in Laṅkā. In a cosmic battle, Rāvaṇa is defeated and Sītā rescued. When Rāma is restored to his kingdom, the populace casts doubt on Sītā's chastity while a captive. To reassure them, Rāma banishes Sītā to a hermitage, where she bears him two sons and eventually dies by reentering the earth from which she had been born. Rāma's reign becomes the prototype of the harmonious and just kingdom, to which all kings should aspire. Rāma and Sītā set the ideal of conjugal love; Rāma's relationship to his father is the ideal of filial love; and Rāma and Lakṣmaṇa represent perfect fraternal love. Everything in the myth is designed for harmony, which after being disrupted is at last regained.

Rāma and Sītā

In all but its oldest form, the *Rāmāyaṇa* identifies Rāma with Vishnu as another incarnation and remains the principal source for Rāmaism (worship of Rāma). Though not as long as the *Mahābhārata*, the text contains a great deal of comparable religious material in the form of myths, stories of great sages, and accounts of exemplary human behaviour.

Rāma also has a shadow side. His killing of the monkey king Vālin (or Bālin) in violation of all rules of combat and his banishment of the innocent Sītā are troublesome to subsequent tradition. These problems of the "subtlety" of dharma and the inevitability of its violation, central themes in both epics, remained the locus of philosophical argument throughout Indian history. Apart from their influence as Sanskrit texts, the *Mahābhārata* and the *Rāmāyaṇa* have made an impact in southern and Southeast Asia, where their stories have been continually retold in vernacular and oral versions, and their influence on Indian and Southeast Asian art has been profound. Even today, the epic stories and tales are part of the early education of all Hindus; a continuous reading of the *Rāmāyaṇa* is an act of great merit, and a popular enactment of one version is an annual event across northern India.

The Bhagavadgītā. The *Bhagavadgītā* ("Song of the Lord") is the most influential Indian religious text, although it is not strictly Sruti, or revelation. It is a brief text, 700 verses divided into 18 chapters, in quasi-dialogue form. When the opposing parties in the *Mahābhārata* war stand ready to begin battle, Arjuna, the hero of the favoured party, despairs at the thought of having to kill his kinsmen and lays down his arms. Krishna, his charioteer, friend, and adviser, thereupon argues against Arjuna's failure to do his duty as a noble. The argument soon becomes elevated into a general discourse on religious and philosophical matters. The text is typical of Hinduism in that it is able to reconcile different viewpoints, however incompatible they seem to be, and yet emerge with an undeniable character of its own. Three different ways of releasing the self from transmigration are set forth. There is the discipline of action (karma-yoga): against the views held by Buddhism, Jainism, and Sāṃkhya philosophy, which hold that all acts bind and that therefore abstention from action is a precondition of release, Krishna argues that it is not the acts that bind but the selfish intentions with which they are performed. He argues for a self-discipline in which a person does his duties according to the dictates of prescribed tasks (dharma), but without any self-interest in the personal consequences of the acts. On the other hand, he does not deny the relevance of the discipline of knowledge (jnana-yoga), in which one seeks release in a yogic (ascetic) course of withdrawal and concentration. Then the tone changes and becomes intensely religious: Krishna reveals himself as the Supreme God and grants Arjuna a vision of himself. The third, and perhaps superior, way of release is through a discipline of devotion to God (bhakti-yoga) in which the self humbly worships the loving God and in release hopes not so much for personal liberation from transmigration but for an eternal vision of God. In response to this devotion, God will extend his grace to his votaries, enabling them to overcome the bonds of this world.

Theologies of the *Bhagavadgītā*

The *Bhagavadgītā* is not a systematic theological treatise, and it combines many different elements from Sāṃkhya and Vedānta philosophy. In matters of religion, its important contribution was the new emphasis placed on devotion, which has since remained a central path in Hinduism. In addition, the popular theism evidenced elsewhere in the *Mahābhārata* and the transcendentalism of the Upanishads converge, and a God of personal characteristics is identified with the *brahman* of the Vedic tradition. In its three disciplines the *Bhagavadgītā* gives a typology of the three dominant trends of Indian religion: dharma-based Brahmanism, enlightenment-based asceticism, and devotion-based theism.

The influence of the *Bhagavadgītā* has been profound. It was a popular text, open to all who would listen, and it was fundamental for all later Hinduism. Vedānta philosophy recognizes it, with the Upanishads and the *Brahmasūtras* (brief doctrinal rules concerning *brahman*), as the third authoritative text, so that all philosophers wrote commentaries on it. Even in the 20th century, as is evident from the lives of such diverse personalities as the Indian freedom fighters Tilak and Gāndhi, who acknowledged its influence, it has continued to shape the attitudes of Hindus.

The *Bhagavadgītā*, by demanding that God's worshipers fulfill their duties—"better one's own duty ill-done than another's well-performed" (3.35)—and observe the rules of moral conduct, bridged the chasm between ascetic disciplines and the search for emancipation, on the one hand, and the exigencies of daily life, on the other. For those who must lead a normal life in this world, the *Bhagavadgītā* gave a moral code and a prospect of final liberation. Thus, the work founded, on the basis of the Vaishnava tradition, what may be called a social ethic. Because God is in all beings as their physical and psychical substratum—and as he exists collectively in human society—the wise should not see any difference between their fellow creatures and should love God in them equally. Like God himself, the devotee should be impartial—the same to friend as to foe. The serious endeavour of realizing God's presence in human beings requires humility and a complete unconsciousness of oneself as a corollary of the consciousness of the Presence. It demands the selfless dedication of all actions, duties, and ceremonies to the Lord and obliges a person to promote both individual and social uplift and welfare. Yet, by emphasizing that all humans have not only different propensities for each of the three disciplines of release but also different responsibilities arising out of their births in different castes, the *Bhagavadgītā* also provided a powerful justification for the caste system.

Popular Sanskrit texts

The Purāṇas. The period of the Guptas saw the production of the first of the series (traditionally 18) of often voluminous texts that treat in encyclopaedic manner the myths, legends, and genealogies of gods, heroes, and saints. The usual list of the *Purāṇas* is as follows: the *Brāhma-*, *Brāhmāṇḍa-*, *Brahmavaivarta-*, *Mārkaṇḍeya-*, *Bhaviṣya-*, and *Vāmana-Purāṇas*; the *Viṣṇu-*, *Bhāgavata-*, *Nāradīya-*, *Gāruḍa-*, *Pādma-*, and *Vārāha-Purāṇas*; and the *Śiva-*, *Liṅga-*, *Skanda-*, *Agni-* (or *Vāyu-*), *Mātsya-*, and *Kūrma-Purāṇas*. Many deal with the same or similar materials.

With the epics, with which they are closely linked in origin, the *Purāṇas* became the scriptures of the common people; they were available to everybody, including women and members of the lowest order of society (Śūdras), and were not, like the Vedas, restricted to initiated

men of the three higher orders. The origin of much of their contents may be non-Brahmanic, but they were accepted and adapted by the Brahmans, who thus brought new elements into their orthodox religion.

At first sight the discontinuity between Vedic and Purāṇic mythology appears to be so sharp that they might be considered as being of altogether different traditions. Yet it soon becomes clear that they are in part continuous and that what appears to be discrepancy is merely a difference between the liturgical emphasis of the Vedas and the more eclectic genres of the epics and *Purāṇas*. For example, the great god of the Rigveda is Indra, the god of war and monsoon, prototype of the warrior; but for the population as a whole he was more important as the rain god than the war god, and it is as such that he survives in early Purāṇic mythology. Little is learned in the Veda of goddesses, yet they rose steadily in recognition in Purāṇic mythology.

Although in the *Purāṇas* some of the Vedic gods have an afterlife in which their importance is reduced, other gods, previously of less official significance, arise. The two principal gods of Purāṇic Hinduism are Vishnu and Rudra-Śiva. Both are known in the Vedas, though they play only minor roles: Vishnu is the strider who, with his three strides, established the three worlds (heaven, atmosphere, and Earth) and thus is present in all three orders; and Rudra-Śiva is a mysterious god who must be propitiated.

Purāṇic literature documents the stages of the rise of the two gods as they eventually attract to themselves the identities of other popular gods and heroes: Vishnu assumes the powers of those gods who protect the world and its order, Śiva the powers that are outside and beyond Vishnu's. To these two is often added Brahmā, creator of the world and teacher of the gods. Although still a cosmic figure, Brahmā appears in the *Purāṇas* primarily to appease overpowerful sages and demons by granting them boons.

In the Purāṇic literature of AD 500 to 1000, sectarianism creeps into mythology, and one god is extolled above the others. Of prime interest are cosmology, myths of the great ascetics (who in some respects eclipse the old gods), and myths of sacred places, usually rivers and fords, whose powers to reward the pilgrim are often cited and related to local legends.

Cosmogony. Purāṇic cosmogony greatly expands upon the already complex cosmogonies of the Brāhmaṇas, Upaniṣads, and epics. According to one of many versions of the story of the origin of the universe, in the beginning the god Nārāyāṇa (identified with Vishnu) floated on the snake Ananta ("Endless") on the primeval waters. From his navel grew a lotus, in which the god Brahmā was born reciting the four Vedas with his four mouths and creating the "Egg of Brahmā," which contains all the worlds. There are numerous other accounts that refer to other demiurges, or creators, like Manu (the primordial ancestor of humankind).

Mythologies of creation and doomsday

Although the Vedas do not seem to conceive of an end to the world, Purāṇic cosmogony accounts for the periodic destruction of the world at the close of an eon, when the Fire of Time will put an end to the universe. Elsewhere the destruction is specifically attributed to the god Śiva, who dances the tandava dance of doomsday and destroys the world. Yet this end is not an absolute end but a temporary suspension (pralaya), after which creation begins again in the same fashion.

Cosmology. The *Purāṇa* texts present an elaborate mythical cosmography. The old tripartite universe persists, but it is modified. There are three levels—heaven, Earth, and the netherworld—but the first and last are further subdivided into vertical layers. Earth consists of seven circular continents, the central one surrounded by the salty ocean and each of the other concentric continents by oceans of other liquids. In the centre of the central mainland stands the cosmic mountain Meru; the southernmost portion of this mainland is Bhāratavarṣa, the old name for India. Above Earth there are seven layers in heaven, at the summit of which is the world of Brahmā (*brahma-loka*); there are also seven layers below Earth, the location of hells inhabited by serpents and demons of various kinds.

Myths of time and eternity. The oldest texts speak little of time and eternity. It is taken for granted that the gods, though born, are immortal; they are called "sons of Immortality." In the Atharvaveda, Time appears personified as creator and ruler of everything. In the Brāhmaṇas and later Vedic texts there are repeated esoteric speculations concerning the year, which is the unit of creation and thus is identified with the creative and regenerative sacrifice and with Prajāpati ("Lord of Creatures"), the god of the sacrifice. Time is an endless repetition of the year, and thus of creation; this is the starting point of later notions of repeated creations.

The four yugas of the world

Purāṇic myths develop around the notion of yuga (world age), of which there are four. These four yugas, Kṛta, Tretā, Dvāpara, and Kali—they are named after the four throws, from best to worst, in a dice game—constitute a *mahāyuga* (large yuga), and, like the comparable ages of the world depicted by the Greek poet Hesiod, are periods of increasing deterioration. Time itself also deteriorates, for the ages are successively shorter. Each yuga is preceded by an intermediate "dawn" and "dusk." The Kṛta Yuga lasts 4,000 years, with a dawn and dusk of 400 years each, or a total of 4,800 years; Tretā a total of 3,600 years; Dvāpara 2,400 years; and Kali (the current one), 1,200 years. A *mahāyuga* thus lasts 12,000 years and observes the usual coefficient of 12, derived from the 12-month year, the unit of creation. These years are "years of the gods," each lasting 360 human years, 360 being the days in a year. Two thousand *mahāyugas* form one kalpa (eon), which is itself but one day in the life of Brahmā, whose full life lasts 100 years; the present is the midpoint of his life. Each kalpa is followed by an equally long period of abeyance (pralaya), in which the universe is asleep. Seemingly the universe will come to an end at the end of Brahmā's life, but Brahmās too are innumerable, and a new universe is reborn with each new Brahmā.

Another myth lays particular stress on the destructive aspect of time. Everything dies in time: "Time ripens the creatures, Time rots them" (*Mahābhārata* 1.1.188). "Time" (*kāla*) is thus another name for the god of death, Yama. The name is associated especially with Śiva in his destructive aspect as Mahākāla and is extended to his consort, who may be known as the goddess Kālī or Mahākālī. On a mythological level the speculations on time reflect the doctrine of the eternal return in the philosophy of transmigration. The universe returns just as, after death, a soul returns to be born again. In the oldest description of the process (Chāndogya Upaniṣhad 5.3.1.–5.3.10), the account is still mythic, but with tendencies to naturalism. The soul on departing may go either of two ways: the Way of the Gods, which brings it through days, bright fortnights, the half year of the northern course of the Sun, to the full year, and eventually to *brahman;* or the Way of the Ancestors, through nights, dark fortnights, the half year of the southern course of the Sun, and, failing to reach the full year, eventually back to Earth clinging to raindrops. If the soul happens to light on a plant that is subsequently eaten by a man, the man may impregnate a woman and thus the soul is reborn. Once more the significance of the year as a symbol of complete time is clear.

Myths of the gods. According to the epic *Mahābhārata* (1.1.39), there are 33,333 Hindu deities. In other, later sources, that number is multiplied a thousandfold. Usually, however, the gods are referred to as "The Thirty-Three."

The tendency toward pantheism increased in Purāṇic Hinduism and led to a kind of theism that exalted several supreme gods who were not prominently represented in the Vedic corpus, while many of the Vedic gods disappeared or were greatly diminished in stature. New patterns became apparent: the notion of rita, the basis of the conception of cosmic order, was reshaped into that of dharma, or the religious-social tasks and obligations of humans in society that maintain order in the universe. There also was a broader vision of the universe and the place of divinity.

Three principal moments are envisioned in the life of the cosmos: creation, maintenance, and destruction. Important myths about the gods are tied to these moments. Traditionally, Brahmā is the creator, emanating the universe and simultaneously promulgating the four Vedas from his four mouths. The conception of time as almost

endlessly repeating itself in kalpas detracts, however, from the uniqueness of the first creation, and Brahmā becomes little more than a demiurge. Far more attention is given to the maintenance and to the destruction of the universe.

Maintenance and destruction are symptomatic of order and disorder, and order and disorder in turn are closely associated with society and the realm outside society. The god Vishnu, who became the god of maintenance, is thus also the social god par excellence, while Śiva, partly established as the agent of destruction, is in many respects an asocial god. Vishnu is the saviour from lawlessness, destroyer of those who threaten the good order, and king of the harmonious realm. Śiva represents untamed wildness; he is the lone hunter and dancer, the yogi (the accomplished practitioner of yoga) withdrawn from society, and the ash-covered ascetic. The distinction between the gods is not between good and evil but rather between two ways in which the divine manifests itself in this world—as both benevolent and fearful, both harmonious and disharmonious.

Bhāgavata-Purāṇa. South Indian devotionalism produced many works in Sanskrit; the most important was the *Bhāgavata-Purāṇa,* which soon became known throughout India. Its 10th book is devoted entirely to Krishna, and there, for the first time, his adventures as a lovable child and as a youth are recounted in great detail. The *Bhāgavata-Purāṇa* may have been written in the 10th century and is certainly a product of the Dravidian south. The doctrine of the avatars of Vishnu was by now in full force, and the *Bhāgavata* recognizes 22 of them.

While all *Purāṇa*s have exerted influence on Hinduism—and are in turn reflections of trends in Hinduism—none can compare in popularity with the *Bhāgavata-Purāṇa* ("The *Purāṇa* of the Devotees of the Blessed Lord Krishna"), the *Purāṇa* of the god Krishna par excellence. It differs from the other *Purāṇa*s in that it is planned as a unit and that far greater care is taken with both metre and style. Its nearly 18,000 stanzas are divided into 12 books.

The childhood of Krishna

The most popular part of the *Purāṇa* is the description of the life of Krishna, for which it has since remained the principal authority. In this work far greater emphasis than in other texts is placed on the youth of Krishna: the threats against his life by the tyrant Kaṃsa, his flight and life among the cowherds at Gokula, and especially his adventures and pranks with the cowherd girls. This treatment has remained classic, and the popularity of the text has led to the survival of many manuscripts, some beautifully illustrated. Much of medieval Indian painting and an enormous amount of vernacular literature draw upon the *Bhāgavata-Purāṇa* for their themes.

The *Bhāgavata-Purāṇa* teaches a quite representative Vaishnava theology: God is transcendent and beyond human understanding; he is the universal causality, creator and substratum; he is time and the bearer of all possibilities that are susceptible of actualization; through his incomprehensible creative ability (maya) or specific power (*ātmaśakti*) he expands himself into the universe, which he pervades and which is his outward appearance (his immanence). Thus he is the All and everything and the inner Self of all beings. When God is conceived of as *brahman,* he is immutable and therefore must be the Puruṣa (cosmic Person) who is not the universe; if, however, his creation is thought to be in him, he is the world.

Accepting the *Bhāgavata-Purāṇa* as a high scriptural authority, Vaishnavism considers God the ground and subsistence of whatever exists, from whom all objects have come, by whom they continue to be, toward whom they move, and into whom they enter at the final dissolution at the end of this world, unless they already came to him in the state of emancipation (moksha). Between God and the world there is a relation of inconceivable difference in identity and identity in difference (*acintyabhedābheda;* literally, "unthinkable difference and nondifference"). The Lord creates the world merely because he wills to do so. Creation, or rather the process of differentiation and integration, is his sport (*līlā*). The world is real, but reality has two aspects: the transcendent and eternally real and the reality that is progressively realized and, in the process, bound up with the eternal aspect.

One of the chief purposes of the *Bhāgavata-Purāṇa* is the glorification of an intensely personal and passionate bhakti that gradually develops into a decidedly erotic mysticism, independent of all alternative means of salvation. According to this text, there are nine characteristics of bhakti: listening to the sin-destroying sacred histories; praising God's name; remembering and meditating on his nature and salutary endeavour (resulting in a spiritual fusion of devotee and God); serving his image; adoring him; respectful salutation; servitude; friendship; and self-surrender. Meritorious works are also an element of bhakti.

Bhagavat and Bhāgavata

According to the *Bhāgavata-Purāṇa,* the highest Bhāgavata—worshiper of the Bhagavat (God: "the Adorable One")—sees himself in all beings and all beings in the Bhagavat; free from hatred and prejudice and knowing God to be present in all beings, he loves him by loving them. Those who cannot reach this level can at least have friendly relations with coreligionists, irrespective of their birth or social status, and take compassion upon the infatuated. The true Vaishnava should worship Vishnu or one of his avatars, construct temples, bathe in holy rivers, study religious texts, serve superiors, and honour cows. In social intercourse with the adherents of other religions he tends to be passively intolerant, avoiding direct contact, without injuring them or prejudicing their rights. He should not neglect other gods but must avoid following the rituals of their followers. Misuse of the advantages of birth is severely condemned, and those who apply themselves mainly to the acquisition and enjoyment of wealth are not well qualified for bhakti. The concept of class divisions is accepted, but the idea that possession of the characteristics of a particular class is the inevitable result of birth is decidedly rejected. Because sin is antithetical to bhakti, a Brahman who is not free from falsehood, hypocrisy, envy, aggression, and pride cannot be the highest of men, and many persons of low social status may have some advantage over him in moral attitude and behaviour. The most desirable behaviour is compatible with bhakti but independent of class.

In establishing bhakti religion against any form of opposition and defending the devout irrespective of birth, the Bhāgavata religion did not actively propagate social reform; but the attempts to make religion an efficient vehicle of new spiritual and social ideas, especially Caitanya's movement, contributed, to a certain extent, to the emancipation of lowborn followers of Vishnu.

VAISHNAVISM AND ŚAIVISM

Vaishnavism. Vaishnavism is the worship of Vishnu and his various incarnations. During a long and complex development from Vedic times, there arose many Vaishnava groups with differing beliefs and aims. Some of the major Vaishnava groups include the Śrīvaiṣṇavas and Dvaitins ("philosophical or religious dualists") of South India, the followers of the teachings of Vallabha in western India, and several Vaishnava groups in Bengal in eastern India, who follow teachings derived from those of the saint Caitanya. The majority of Vaishnava believers, however, take what they like from the various traditions and blend it with various local practices.

In the Veda, Vishnu is the god of far-extending motion and pervasiveness who, for humans in distress, particularly through constrictions, penetrates and traverses the triple spaces to make their existence possible. All beings are said to dwell in his three strides or footsteps (*tri-vikrama*): his highest step, or abode, is beyond mortal ken in his dear and highest resort, the realm of heaven. So Vishnu is also the god of the pillar of the universe and is identified with the sacrifice. He imparts his all-pervading power to the sacrificer who imitates his strides and so identifies himself with the god, thus conquering the universe and attaining "the goal, the safe foundation, the highest light" (Śatapatha Brāhmaṇa).

Composite historical nature of Vishnu

In the centuries preceding the beginning of the Christian era, Vishnu became the Īśvara (immanent deity) of his special worshipers, fusing with the Puruṣa-Prajāpati figure; with Nārāyaṇa, whose cult discloses a prominent influence of ascetics; with Krishna, who in the *Bhagavadgītā* revealed a popular and universal religion, open to every-

body desiring to lead a socially normal life while having a prospect of final liberation; and with Vāsudeva, adored by a group known as the Pāñcarātras.

The extensive mythology attached to Vishnu consists largely of the mythology of his incarnations (avatars). Although the notion of "incarnation" is found elsewhere in Hinduism, it is basic to Vaishnavism. The concept is particularly geared to the social role of Vishnu; whenever dharma (universal law and order) is in danger, Vishnu departs from his heaven, Vaikuṇṭha, and incarnates himself in an earthly form to restore the good order. Each of his incarnations has a particular mythology. The classical number of these incarnations is 10, ascending from theriomorphic (animal form) to fully anthropomorphic manifestations. These are: Fish (Matsya), Tortoise (Kūrma), Boar (Varāha), Man-Lion (Narasiṃha), Dwarf (Vāmana), Rāma-with-the-Ax (Paraśurāma), King Rāma, Krishna, Buddha, and the future incarnation, Kalkin.

The 10 avatars of Vishnu

A god thus active for the good of society and the individual inspires love. Vishnu has indeed been the object of devotional religion (bhakti) to a marked degree, but mainly in his incarnations, and among them specially as Krishna and Rāma. The god rewards devotion with his grace, through which the votary may be lifted from transmigration to release. Like most other gods, Vishnu has his especial entourage: his wife is Lakṣmī or Śrī, the lotus goddess, granter of beauty, wealth, and good luck. She came forth from the ocean when gods and demons churned it in order to recover from its depths the ambrosia or elixir of immortality, *amṛta*. At the beginning of the commercial year special worship is paid to her for success in personal affairs. Vishnu's mount is the bird Garuḍa, archenemy of snakes, and his emblems are the lotus, club, discus (as a weapon), and a conch shell, which he carries in his four hands.

Whatever justification the different Vaishnava groups (*e.g.*, the Śrīvaiṣṇavas of South India or the worshipers of Vishnu Viṭhobā in Mahārāshṭra) offer for their philosophical position, all Vaishnavas believe in God as a person with distinctively high qualities and worship him through his manifestations and representations. Vaishnava faith is essentially monotheistic, whether the object of adoration be Vishnu Nārāyaṇa or one of his avatars such as Rāma or Krishna. Preference for any one of these manifestations is largely a matter of tradition. Thus, most South Indian Śrīvaiṣṇavas prefer Vishnu, Rāma, or Śrī (Vishnu's consort); the North Indian groups prefer Krishna. The avatar doctrine, by accommodating the cults of various divine or heroic figures within a monotheistic framework, proved to be a powerful integrating force. Whenever the dharma declines and evil and general disaster threaten, God, the protector and preserver of the world, emanates himself and assumes an earthly form to guard the good, to destroy the wicked, and to confirm the dharma. The benevolence and beneficial activity of these figures (Rāma, Krishna, *et al.*) is, however, occasionally in doubt. In many mythical tales, Vishnu is depicted as a versatile figure of great adaptability, able, for instance, to disguise himself as a fascinating young woman in order to trick the asuras (antigods) out of the possession of the newly produced *amṛta*. His absorbing, many-sided character was a source of inspiration for various stories in which he often acts deceitfully, selfishly, or helplessly. The scene of his great deeds is usually laid in this world, especially India, in places often mentioned by name. The narratives are full of the miraculous, but their central figures give the impression of human, sometimes all too human, characters whose actions and reactions are within the limits of ordinary understanding.

The playfulness of the incarnate God

A pronounced feature of Vaishnavism is the strong tendency to devotion (bhakti), which is generally considered to be "the heart of worship," the sole true religious attitude toward a personal God, and the very foundation of the realization of man's relationship with him. Characterized by a continual consciousness of participating in God's essence, bhakti is the disinterested performance of all deeds for him, a passionate love and adoration of God, and a complete surrender to him. The widespread bhakti movement is a corollary of the Vaishnava ideal of a loving personal God and aversion to a conception of salvation that puts an end to all consciousness or individuality. Attesting to the superiority of a mystic and emotional attitude to the meditative or preponderantly ritualistic means to the highest goal, the practical and theoretical development of the bhakti idea constitutes one of the main points of difference among the several Vaishnava schools. The belief expressed in the *Bhagavadgītā*—that those who seek refuge in God with all their being will, by his benevolence and grace (*prasāda*), win peace supreme, the eternal abode—was generally accepted: bhakti will result in divine intercession with regard to the consequences of one's deeds. Among many followers of Rāmānuja, however, complete self-surrender (*prapatti*) came to be distinguished from bhakti as a superior means of spiritual realization.

Śaivism. The character and position of the Vedic god Rudra—called Śiva, "the Mild or Auspicious One," when this aspect of his ambivalent nature is emphasized—remain clearly perceptible in some of the important features of the great god Śiva, who together with Vishnu came to dominate Hinduism.

During a complex development from ancient, possibly in part from pre-Vedic, times, many different Śaiva groups arose. Major groups such as the Liṅgāyats of southern India and the Kashmir Śaivas contributed the theological principles of Śaivism, and Śaiva worship became a complex amalgam of pan-Indian Śaiva philosophy and local or folk worship.

In the minds of the ancient Indians Śiva must have been primarily the divine representative of the uncultivated, dangerous, unreliable, and much-to-be-feared aspects of nature. Śiva's character lent itself to being split into partial manifestations—each said to represent only an aspect of him—as well as to assimilating divine or demoniac powers of a similar nature from other deities. Already in the Rigveda, appeals to him for help in case of disaster—of which he might be the originator—were combined with the confirmation of his great power. In the course of the Vedic period, Śiva—originally a ritual and conceptual outsider, yet a mighty god whose benevolent aspects were readily emphasized—gradually gained access to the circle of respectable gods who preside over various spheres of human interest. Many characteristics of the Vedic Prajāpati, the creator, of Indra, the god of the phallus, and of the great Vedic god of fire, Agni, have been integrated into the figure of Śiva.

Partial manifestations of Śiva

In those circles that produced the Śvetāśvatara Upanishad (*c.* 400 BC), Śiva rose to the highest rank. Its author uses grandiose terms to show a way of escape from samsara, to proclaim Śiva the sole eternal Lord, and to establish Śiva's existence. In this description of Śiva's nature, some of the most salient features of the later Śiva, the Īśvara (immanent deity), are clearly discernible: he is the ultimate foundation of all existence and the source and ruler of all life, who, while emanating and withdrawing the universe, is the goal of that identificatory meditation that leads to complete cessation from phenomenal existence. While Vishnu became a friend nearer to man, Rudra-Śiva developed into an ambivalent and many-sided lord and master. His "doubles" or partial manifestations, however, were active among mankind: as Paśupati ("Lord of Cattle"), he took over the fetters of the Vedic Varuṇa; as Aghora ("To Whom Nothing Is Horrible"), he showed the uncanny traits of his nature (evil, death, punishment) and also their opposites.

It is not always clear in particular cases whether Śiva is invoked as a great deva (god) of frightful aspect, capable of conquering demoniac power, or as the boon-giving Lord and protector. The Īśvara idea of a Highest Being demonstrably beyond contingency is rather abstract; hence its propagators needed to use imagery, popular belief, and mythical thought. Śiva might be the sole Principle above change and variation, yet he did not sever his connections with innumerable local deities and much-feared powers worshiped by most Hindus, who still continue to invoke him in magical rites. Whereas Vishnu champions the cause of the gods, Śiva sometimes sides with the demons.

Śiva is a typical example of polarity within the Highest Being because he reconciles in his person semantically opposite though complementary aspects: he is both terri-

ble and mild, creator and agent of reabsorption, eternal rest and ceaseless activity. These seeming contradictions make him a paradoxical figure, transcending humanity and assuming a mysterious sublimity of his own. His character is so complicated and his interests are so widely divergent as to lead him in mythical narratives into conflicting situations. Yet, although Brahman philosophers like to emphasize his ascetic aspects and the ritualists of the Tantric tradition his sexuality, the seemingly opposite strands of his nature are generally accepted as two sides of one character.

Power of chastity and sexuality

Śiva interrupts his austerity and asceticism (*tapas*), which is sometimes described as continuous, to marry Pārvati—he is even said to perform ascetic acts in order to win her love—and he combines the roles of lover and ascetic to such a degree that his wife must be an ascetic (yogi) when he devotes himself to austerities and a lustful mistress when he is in his erotic mode. This dual character finds its explanation in the ancient double conviction that unrestrained sexual intercourse is conducive to the fertility of nature and that the chastity and continence of the ascetic produce marvelous events and have an uncommon influence upon the unseen. By his very chastity, an ascetic accumulates (sexual) power that can be discharged suddenly and completely so as to produce marvelous results such as the fecundation of the soil. From various mythical tales it is seen that both chastity and the loss of chastity are necessary for fertility and the intermittent process of regeneration in nature. Ascetics engaging in erotic and creative experiences are a familiar feature in Hinduism, and the elements of teeming sexuality in mythological thought counterbalances the Hindu bent for asceticism. Such sexuality, while rather idyllic in Krishna, assumes a mystical aspect in Śiva, which is why the devotee can see in him the realization of the possibilities of both asceticism and the householder state. His marriage with Pārvati is, then, a model of conjugal love, the divine prototype of human marriage, sanctifying the forces that carry on the human race.

Śiva's myths tend to depict him as the absolutely mighty unique One, who is not responsible to anybody or for anything. His many poses express aspects of his nature: as a dancer, he is the originator of the eternal rhythm of the universe; he also catches the waters of the heavenly Ganges River, which destroy all sin; and he wears in his headdress the crescent moon, which drips the nectar of everlasting life.

Śiva represents the unpredictability of divinity. In him the Vedic Rudra is partly continued, but his mythology has become exceedingly complex. He is the hunter who slays and skins his prey and dances a wild dance while covered with the bloody hide. Far from society and the ordered world, he sits on the inaccessible Himalayan plateau of Mount Kailāsa, an austere ascetic, averse to love, who burns Kāma, the god of love, to ashes with a glance from the third eye—the eye of insight beyond duality—in the middle of his forehead. Yet another epiphany is that of the lingam, an upright rounded post, usually of stone, a formalized phallic symbol, in which form he is worshiped throughout India. And at the end of the eon, he will dance the universe to destruction. He is, nevertheless, invoked as Śiva, Śambhu, Śaṅkara (meaning: "the Auspicious One" or "the Peaceful One"), for the god that can strike down can also spare. Snakes seek his company and twine themselves around his body. He wears a necklace of skulls. He sits in meditation, with his hair braided like a hermit's, his body smeared white with ashes. These ashes recall the burning pyres on which the sannyasis (renouncers) take leave of the social order of the world and set out on a lonely course toward release, carrying with them a human skull.

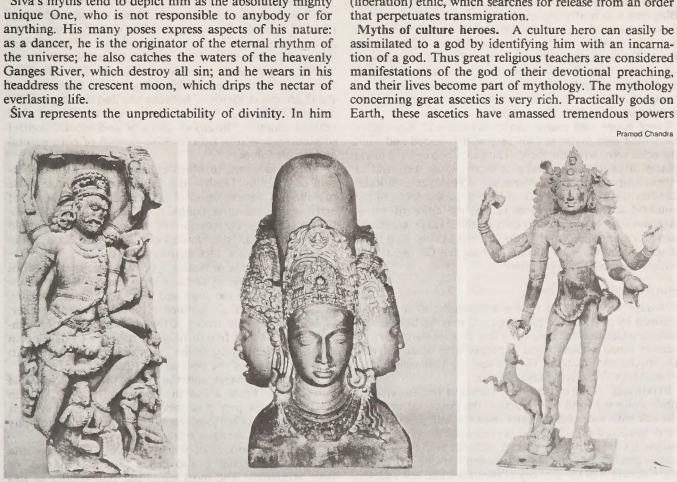
Śiva and his consorts

Like so many ascetics—often irascible and dangerous—Śiva demands to be seduced. His consort is Pārvati ("Daughter of the Mountain"), a goddess most unlike the consorts of Vishnu in his various incarnations. She is also personified as the Goddess (Devī "goddess"), Mother (Ambā), black and destructive (Kālī), fierce (Caṇḍikā), and well-nigh inaccessible (Durgā). As Śiva's female counterpart, she inherits some of Śiva's more fearful aspects. She comes to be regarded as the power (śakti) of Śiva, without which Śiva is literally powerless. Śakti is in turn personified in the form of many different goddesses, often said to be aspects of her.

Thus the spheres of the Vishnu complex and the Śiva complex are very different ones. In important respects they represent the two different ethics of Hinduism: the dharma ethic, which aims at upholding the dharma and the cosmic and social order based on it, and the moksha (liberation) ethic, which searches for release from an order that perpetuates transmigration.

Myths of culture heroes. A culture hero can easily be assimilated to a god by identifying him with an incarnation of a god. Thus great religious teachers are considered manifestations of the god of their devotional preaching, and their lives become part of mythology. The mythology concerning great ascetics is very rich. Practically gods on Earth, these ascetics have amassed tremendous powers

Pramod Chandra



Paradoxical nature of Śiva as indicated by differing representations.
(Left) Gajāntaka, destroyer of the elephant demon, sculpture from Orissa, 8th century. In the Indian Museum, Calcutta. (Centre) Caturmukha Linga, lingam or phallic symbol, with four faces, sculpture, 5th century. In worship at Nācna-Kuṭhārā, Madhya Pradesh, India. (Right) Bhikshatana, the naked ascetic with his begging bowl, early Cōla bronze, from Tiruvengadu District, 1048. In the Thanjāvūr Art Gallery, India.

that they do not hesitate to use. The sage Kapila, meditating in the netherworld, burned to ashes 60,000 princes who had dug their way to him. Another sage, Bhagiratha, brought the Ganges River down from heaven to sanctify their ashes and, in the process, created the ocean. Agastya, revered as the Brahman who brought Sanskrit civilization to South India, drank and digested the ocean. When the Vindhya mountain range would not stop growing, Agastya crossed it to the south and commanded it to cease growing until his return; he still has not returned. Viśvāmitra, a king who became a Brahman, created a new universe with its own galaxies to spite the gods. It is in such myths that the mythopoeic imagination exults in its sensitivity to the awesome, mysterious, and marvelous.

The trial of kings

In myths concerning kings and princes, a prevailing theme is the trial of the son by the father. For example, the ancient king Yayāti had five sons to whom he wanted to transfer his own senescence for a stipulated period. All refused except the youngest, Puru. As a reward he became his father's successor, and his descendants became the Pauravas, the line of succession or dynasty in which the heroes of the *Mahābhārata* were later born. The latter heroes also underwent a trial when they were exiled from their newly won kingdom; similarly, Rāma underwent his ordeal in exile. Heroines undergo their own trials, which usually challenge their chastity, as in the case of Sītā in the *Rāmāyaṇa* and Draupadī, the one wife of all five Pāṇḍava brothers, whose sari became endless when a lustful villain attempted to pull it off.

Moving from myth to legend, there are also stories told of the great teachers, and every founder of a sect is soon deified as an incarnation of a god: the philosopher Śaṅkara (*c.* 788–820) as an incarnation of Śiva, the religious leader Rāmānuja (d. AD 1137) as that of Nārāyaṇa-Vishnu, and the Bengal teacher Caitanya (1485–1533) simultaneously as that of Krishna and his beloved Rādhā.

Myths of holy rivers and places. Of particular sanctity in India are the perennial rivers, among which the Ganges stands first. This river, personified as a goddess, originally flowed only in heaven until she was brought down by Bhagīratha to purify the ashes of his ancestors. She came down reluctantly, cascading first on the head of Śiva, in order to break her fall, which would have shattered the Earth. Confluences are particularly holy, and the Ganges' confluence with the Yamuna at Allahābād is the most sacred spot in India. Another river of importance is the Sarasvatī, which loses itself in desert; it was personified as a goddess of eloquence and learning.

Every major and many minor temples and sanctuaries have their own myths of how they were founded and what miracles were wrought there. The same is true of famous places of pilgrimage, usually at sacred spots near and in rivers; important among these are Vrindāvana (Brindāban) on the Yamuna, which is held to be the scene of the youthful adventures of Krishna and the cowherd wives. Another such centre with its own myths is Gayā, especially sacred for the funerary rites that are held there. And there is no spot in Vārānasī (Benares) along the Ganges that is without its own mythical history.

PHILOSOPHICAL TEXTS

Although the details of Indian philosophy, as it was developed by professional philosophers, may be treated as a subject separate from Hinduism (see INDIAN PHILOSOPHY), certain broad philosophical concepts were absorbed into the myths and rituals of Hindus and are best viewed as a component of the religious tradition.

Mysticism. One of the major trends of Indian religious philosophy is a kind of mysticism: the desire for union of the self with something greater than the self, whether that be defined as a principle that pervades the universe or as a personal God. Hindu mysticism includes both these forms and a great many that lie in between. At one extreme is the realization of the identity of the individual self with the impersonal principle called *brahman*, the position of the Vedānta school of Indian philosophy; and at the other is the intensive devotionalism to a personal God, called by a variety of names, that is found in the bhakti (devotional) sects.

There are four things common to most Hindu mystical thought. First, it is based on experience: the state of realization, whatever it is called, is both knowable and communicable, and the systems are all designed to teach people how to reach it. It is not, in other words, pure speculation. Second, it has as its goal the release of the spirit-substance of the individual from its prison in matter, whether matter be considered real or illusory. Matter is the cause of the suffering of which Buddhism speaks. Third, all the systems recognize the importance or the necessity of the control of the mind and body as a means of realization; sometimes this takes the form of extreme asceticism and mortification, and sometimes, at the other extreme, it takes the form of the cultivation of mind and body in order that their energies may be properly channeled. And, finally, at the core of Hindu mystical thought is the functional principle that knowing is being. Thus, knowledge is something more than analytical categorizing: it is total understanding. This understanding can be purely intellectual, and some schools equate the final goal with omniscience, as does yoga. Knowing can also mean total transformation: if one truly knows something, he is that thing. Thus, in the devotional schools, the goal of the devotee is to transform himself into a being who, in eternity, is in immediate and loving relationship to the deity. But despite the fact that these are both ways of knowing, the difference between them is significant. In the first instance, the individual has the responsibility to train and use his own intellect. The love relationship of the second, on the other hand, is one of dependence, and the deity assists the devotee through grace. The distinction is generally made by the analogy of the cat and the monkey: the cat carries her young in her mouth, and thus the kitten has no responsibility. But the young monkey must cling by its own strength to its mother's back.

Knowledge as being and becoming

It is usual for writers on the subject, following Surendranath Dasgupta, a historian of Indian philosophy, to list five major varieties of Hindu mysticism, the five having arisen in historical order as follows:

1. The sacrificial, based on the Vedas and Brāhmaṇas.
2. The Upanishadic, in which are found the beginnings of both monistic (concerned with a unitary principle of reality, immanent in the world) and theistic (concerned with a personal or suprapersonal God) systems.
3. The yogic, relating to physical and mental discipline; the earliest known text of this school is the *Yoga-sūtra* of Patañjali, dated variously between the 2nd century BC and the 5th century AD. According to yogic mysticism, man realizes union by means of physical and mental control of himself, which in turn leads to control of both natural and divine forces.
4. The Buddhistic, in which enlightenment is the realization of the four Truths—the fact of suffering, the cause of suffering, the cessation of suffering, and the means of arriving at these three truths: the Eightfold Path. The ultimate state, the culmination of one path of the Eightfold Path, is nirvana, "the blowing out," the extinction of desire (see BUDDHISM, THE BUDDHA AND).
5. The devotional, or bhakti, type of mysticism comprises a range of theistic systems, with a conception of absolute dualism between man and God on the one extreme, and a conception of qualified nondualism on the other. Although there are traces of this devotionalism throughout the history of Indian religion, it began to develop in earnest in South India in the 7th through 10th century AD with the hymns of the poet saints called Āḻvārs.

Philosophical sutras and the rise of the six schools of philosophy. From about the beginning of the Christian era through the period of the Gupta empire, the systems of the Six Schools (Ṣaḍdarśana) of orthodox philosophy were formulated in terse sutras.

Śaṅkara and the Vedānta school

The most important of the Six Schools is the Vedānta ("End of the Vedas"), also called Uttara-Mīmāṃsa, or later Mīmāṃsā. The most renowned philosopher of this school and, indeed, of all Hinduism was Śaṅkara (traditionally dated *c.* 788–820, but he probably died about 20 years later). He was born at Kālaḍi in Kerala and is said to have spent most of his life traveling through India debating with members of other sects. The Śaṅkaran system

has sounded the keynote of intellectual Hinduism down to the present, but later teachers founded sub-schools of Vedānta, which are perhaps equally important.

Śaṅkara was also responsible for the growth of Hindu monasticism, which had been in existence for more than a millennium in the form of hermit colonies. Inscriptions from Gupta times onward also refer to monklike orders of Śaiva ascetics, apparently living according to distinctive disciplines and with distinguishing garments and emblems. Śaṅkara founded a closely disciplined Śaiva order, perhaps partly modeled on the Buddhist sangha, or order, known as *daśnāmī*, which is still the most influential orthodox Hindu ascetic group. The order is composed of 10 brotherhoods and hence called *daśnāmīs* ("those with 10 names"). Orders became an established institution with wider geographic affiliations. Some of these admitted Brahmans only; others were open to all four classes or even to women; some made a practice of nudity. These Śaiva communities are more inclined to individual asceticism and are less closely organized than the Vaishnava *vairāgiṃs* ("the dispassionate") or *gosvāmiṃs* ("the masters"). Śaṅkara is also said to have founded the four main monasteries (*maṭha*) at the four corners of India: Sringeri in Karnataka, Badrināth in the Himalayas, Dwārkā in Gujarāt, and Puri in Orissa. The abbots of these monasteries control the spiritual lives of many millions of devout Śaiva laymen throughout India, and their establishments strive to maintain the traditional philosophical Hinduism of the strict Vedānta. In modern times, certain *daśnāmī* leaders have incurred criticism for their firm opposition to social change.

Vaishnava theology of Rāmānuja

The theologians had to assume the task of explaining the relation between God, as the unaffected and unchanging cause of all things, and the universe. According to Rāmānuja (*c.* 1017–1137), a great South Indian thinker of Śrīvaiṣṇava persuasion, *brahman* (*i.e.*, God) is a Person with high attributes, the object of an individual's search for the higher knowledge that is the only entrance to salvation. Because an absolute creation is denied, God is viewed as the sole cause of his own modifications; namely, the emanation, existence, and absorption of the universe. Although unlimitedly expansive, God is conceived to be essentially different from everything material, the absolute opposite of any evil, free from any imperfection, omniscient, omnipotent, possessed of all positive qualities (such as knowledge, bliss, beauty, and truth), of incomparable majesty, the inner soul of all beings, and the ultimate goal of every religious effort. The universe is considered a real transformation of *brahman*, whose "body" consists of the conscious souls and everything unconscious in their subtle and gross states. The karma doctrine is modified as follows: the Lord, having determined good and bad deeds, provides all individual souls with a body in which they perform deeds, reveals to them the scriptures from which they may learn the dharma, and enters into them as their internal regulator. The individual acts at his own discretion but needs the Lord's assent. If the devotee wishes to please him, God induces him, with infallible justice and loving regard, to intentions and effort to perform good deeds by which the devotee will attain to him; if not, God keeps him from that goal.

The influence of the bhakti movement had earlier led Rāmānuja to admit a twofold possibility of emancipation: in addition to the meditative method of the highest insight (jnana) into the oneness of soul and God, which destroys the residues of karma and propitiates God to win his grace, there is the way of bhakti. Those who prefer the former way will reach a state of isolation, the others an infinitely blissful eternal life in, through, and for God, with whom they are one in nature but not identical. They do not lose their individuality and may even meet Vishnu in his Vaikuṇṭha heaven and enjoy delight beyond description.

An interesting development of Rāmānuja's doctrine of qualified monism is found in the philosophy of Madhva (died *c.* 1276). This Kanarese Brahman taught a doctrine of dualism according to which God and the soul are eternally distinct.

Authors of Śaiva *Purāṇas* established two ingenious and complementary doctrines to explain the nature and omnipotence of God (the force that rules, absorbs, and reproduces the world and that in performing any one of these acts necessarily performs the other two as well), the existence of the world, and the identity of God and the world. A theory of five "faces," or manifestations—each of which is given mythological names and related mantras—is of great ritual significance. It associates Śiva's so-called creative function, by which he provokes the evolution of the material cause of the universe, with his first face, or aspect; its maintenance and reabsorption with his second and third faces; his power of obscuration, by which he conceals the souls in the phenomena of samsara, with his fourth face; and his ability to bestow his grace, which leads to final emancipation, with his fifth face. The five functions are an emanation of the unmanifested Śiva who is the transcendent *brahman*.

The faces became the central elements of a comprehensive classification system. They were identified with parts of God's body, regions of the universe, various ontological principles, organs of sense and action, and the elements. The system was used to explain how Śiva's being is the All and how the universe is exclusively composed of aspects or manifestations of Śiva. In his fivefold nature, Śiva was shown to be identical with the 25 (five times five) elements or principles assumed by the prominent Saṃkhyā school of Indian philosophy. The special significance of the number five in Śaivism can be understood as a philosophical elaboration of the time-honoured fourfold organization of the universe. (The four quarters of the sky also play a prominent part in religious practice.) According to this conception, a fifth aspect, when added to the four, is considered the most important aspect of the group because it represents each of the four and collectively unites all of their functions in itself. The system finds its complement in the doctrine of the five Sadākhyas (five items that bear the name sat, "is" or "being") representing the five aspects of that state, which may be spoken of as the experience of "there is" (sat) and which have evolved from God's fivefold creative energy (shakti). In these, God "dwells" in his aspect called Sadāśiva ("the Eternal Śiva"), which is regarded sometimes as a manifestation of and sometimes as identical with the Supreme Being.

Another Śaiva doctrine posits eight "embodiments" of Śiva as the elements of nature (ether, wind, fire, water, earth), Sun, Moon, and the sacrificer, or consecrated worshiper (also called Atman). To each of these eight elements corresponds one of Śiva's traditional names or aspects—to the last one, usually Paśupati. The world is a product of these eight forms, consists of them, and can only exist and fulfill its task because the eight embodiments cooperate. Because each individual is also composed of the same eight realities (*e.g.*, the light of man's eyes corresponds to that of the Sun), Śiva constitutes the corporeal frame and the psychical organism of every living being. The eighth constituent is the indispensable performer of the rites that sustain the gods who preside over the cosmic processes and are really Śiva's faculties.

Regional variations in Śaivism

Although Śaivism is a much more coherent whole than Vaishnavism, there evolved, in different parts of India, some branches with peculiarities of their own. According to the pronouncedly idealist monism of Kashmir Śaivism, an important religiophilosophic school, Śiva manifests himself through a special power as the first cause of creation, and he also manifests himself through a second power as the innumerable individual souls who, because of a veil of impurity, forget that they are the embodiment of the Highest. This veil can be torn off by intense faith and constant meditation on God, by which the soul transmutes itself into a universal soul and eventually attains liberation through a lightninglike, intuitive insight into its own nature. Those Hindus who adhere to this group consider their doctrine a manifestation of the highest Reality, Knowing Consciousness, neither personal nor impersonal; as Śiva in the form of the transcendent Word, which is his unspoken Thought, the content of which is the universe.

The Śaiva-siddhānta, a prominent religiophilosophic school of Tamil-speaking South India, assumes three eternal principles: God (who is independent existence, unqualified intelligence, and absolute bliss), the universe, and the

souls. The world, because it is created by God (efficient cause) through his conscious power (instrumental cause) and maya (material cause), is no illusion. The main purpose of its creation is the liberation of the beginningless souls, which are conceived as "cattle" (*paśu*) bound by the noose (*pāśa*) of impurity (*mala*) or spiritual ignorance, which forces them to produce karma. However, they see the karma process as a benefit, for as soon as the soul has sufficiently ripened and reached a state of purity enabling it to strive after the highest insight, God graciously intervenes, appearing in the shape of a fully qualified and liberated spiritual guide (guru), through whose words God permits himself to be realized by the individual soul.

TANTRISM

Tantric traditions and Śāktism. Toward the end of the 5th century, the cult of the Mother Goddess began to achieve a significant place in religious life. Śāktism, the worship of the Śakti, the active power of the godhead conceived in feminine terms, should be distinguished from Tantrism, the search for spiritual power and ultimate release by means of the repetition of sacred syllables and phrases (mantras), symbolic drawings (mandalas), and other secret rites elaborated in the texts known as tantras ("looms").

In many respects the tantras are similar to the *Purāṇas*. Theoretically a tantra deals with (1) knowledge, or philosophy, (2) yoga, or concentration techniques, (3) ritual, which includes the formation of icons and the building of temples, and (4) conduct in religious worship and social practice. In general the last two subjects preponderate, while yoga tends to centre on the mystique of certain sound-symbols (mantras) that sum up esoteric doctrines. The philosophy tends to be a syncretistic mixture of Sāṃkhya and Vedānta philosophical thought, with special and at times exclusive emphasis on the god's power, or shakti. The Tantric texts can be divided into three classes: (1) Śaiva *Āgamas* (traditions of the followers of Śiva), (2) Vaishnava *Saṃhitās* ("Collections of the Vaishnavas," a name borrowed from the Vedic Saṃhitās), and (3) Śākta Tantras ("Looms of the Followers of the Goddess Śakti"). However, they all have the common bond of venerating the Goddess.

Śaiva, Vaishnava, and Śākta tantras

Surviving Hindu tantras were written much later than many of those of Tantric Buddhism, and it may be that the Hindus derived much from the Buddhists in this respect. Although there is early evidence of Tantrism and Śāktism in other parts of India, the chief centres of both were modern Bengal, Bihār, and Assam.

Śaiva Āgamas. Like much other Hindu sacred literature, this literature is neither well-cataloged nor thoroughly studied. It is only possible here to summarize classes of texts within the various traditions.

The sects of Āgamic Śaivas (Śiva worshipers who follow their own Āgama—"traditional"—texts) encompass both the Sanskritic Śaiva-siddhānta—*i.e.*, those who accept the philosophical premises and conclusions of Śaivas in the north—and the southern Liṅgāyats or Vīraśaivas (from *vīra*, literally "hero"; a liṅgam is the Śiva emblem that is worshiped in lieu of images). The Śaiva-siddhānta traditionally has 28 *Āgamas* and 150 sub-*Āgamas*. Their principal texts are hard to date; most probably they do not antedate the 8th century. Their doctrine states that Śiva is the conscious principle of the universe, while matter is unconscious. Śiva's power, or shakti, personified as a goddess, causes bondage and release. She is also the magic Word, and thus her nature can be sought out and meditated upon in mantras.

Kashmir Śaivism begins with the *Śivasūtra* or "Lines of Doctrine Concerning Śiva" (*c.* 850) as a new revelation of Śiva. The system embraces the *Śivadṛṣṭi* ("A Vision of Śiva") of Somānanda (950), in which emphasis is placed on the continuous recognition of Śiva; the world is a manifestation of Śiva brought about by his shakti. The system is called *trika* ("triad"), because it recognizes the three principles of Śiva, Śakti, and the individual soul. Vīraśaiva texts begin at about 1150 with the *Vācanam*s ("Sayings") of Basava. The sect is puritanical, worships Śiva exclusively, rejects the caste system in favour of its own social organization, and is highly structured with monasteries and gurus.

Vaishnava Saṃhitās. These consist of two groups of texts: *Vaikhānasa Saṃhitā*s and *Pāñcarātra Saṃhitā*s. The latter group is the prevailing one; more than 200 titles are known, though the official number is 108. *Vaikhānasa Saṃhitā*s (collections of the Vaishnava school of *Vaikhānasa*s, who were originally ascetics) seem to have embodied the original temple manuals for the Bhāgavatas (devotees), which by the 11th or 12th century had become supplanted by the *Pāñcarātra Saṃhitā*s (collections of the Vaishnava school of Pāñcarātra—"the System of the Five Nights"). The philosophy of the latter is largely a matter of cosmogony, greatly inspired by both the Sāṃkhya and yoga philosophies.

Notion of shakti in Vaishnavism

Apart from their theology, in which for the first time the notion of shakti is introduced into Vaishnavism, they are important because they give an exposition of Vaishnava temple and cult practices. On the philosophical side it is maintained that the supreme god Krishna Vāsudeva manifests himself in four coequal "divisions" (*vyūha*s), representing levels in creation. These gods emanate as supramundane patrons before the primary creation is started by their shakti (power). In the primary creation Śakti manifests herself as a female creative force inspired by the Sāṃkhya philosophy's cosmogony. Practically, stress is laid on a type of incarnation—"iconic incarnation"—in which the god is actually present with a portion of himself in a stone or statue, which thus becomes an icon; therefore the icon can be worshiped as God himself.

Śākta Tantras. Śāktism in one form or the other has been known since Bāṇa (*c.* 650) wrote his *Hundred Couplets to Caṇḍī* (*Caṇḍī-śataka*) and Bhavabhūti his play *Mālatī Mādhava* (725), both of which refer to Tantric practices. There is no traditional authoritative list of texts, but many texts are extant.

Śāktism is an amalgam of Śaivism and folk mother-goddess cults. The Śaiva notion that not Śiva himself but his shakti (sexual, creative power) is active is taken to the extreme—that, without Śakti, Śiva is a corpse, and simultaneously that Śakti is the creator as well as creation. In yoga, great importance is ascribed to mantras, which conjure up the realities with which they are identified. Another important notion (partly derived from yoga philosophy) is that through the body run subtle canals that carry esoteric powers connected with the spinal cord, at the bottom of which the Goddess is coiled around the liṅgam as Kuṇḍalinī; she can be made to rise through the body to the top, whereupon release from samsara takes place. Important among the Śākta Tantras are the *Kulārṇava Tantra* ("Ocean of Tantrism"), which gives details on the "left-handed" cult forms of ritual copulation; the *Kulacūḍāmaṇi* ("Crown Jewel of Tantrism"), which embroiders on ritual; and the *Śāradātilaka* ("Beauty Mark of the Goddess Śāradā") of Lakṣmaṇadeśika (11th century), which discusses magic.

A temple was erected in honour of the mother goddesses at Gaṅgdhār, Rājasthān, in AD 423. There, magical rites of a terrifying kind were practiced, for the temple is described as "loud with the shouts of demonesses, crying in the thick darkness," by the playwright Bhavabhūti, whose drama *Mālatī Mādhava* (about the adventures of the hero Mādhava and his beloved Mālatī) contains a scene depicting secret rites with human sacrifice and ritual cannibalism. The goddess cults eventually centred around Durgā, the consort of Śiva, in her fiercer aspect.

Nature of Tantric tradition. Tantrism, which appears both in Buddhism and in Hinduism, is an important component of religion that, though primarily meant for esoteric circles, also influenced, from the 5th century AD, many religious trends and movements. Opinions of what Tantrism is are quite diverse. Generally, Tantrism claims to show in times of religious decadence a new way to the highest goal and bases itself upon mystic speculations concerning divine creative energy (shakti). Tantrism is a method of conquering transcendent powers and realizing oneness with the highest principle by yogic and ritual means—in part magical and orgiastic—which are also supposed to achieve other supranormal goals.

Tantrists take for granted that all factors in both the macrocosm and the microcosm are closely connected. The adept (*sādhaka*) has to perform the relevant rites on his own body, transforming its normal, chaotic state into a "cosmos." The macrocosm is conceived as a complex system of powers that by means of ritual-psychological techniques can be activated and organized within the individual body of the adept. Contrary to the ascetic emancipation methods of other groups, the Tantrists emphasize the activation and sublimation of the possibilities of their own body, without which salvation is believed to be beyond reach.

Vāmācāra: left-handed Tantra

The Tantrists of the *Vāmācāra* ("the left-hand practice") sought to intensify their own sense impressions by making enjoyment, or sensuality (*bhoga*), their principal concern: the adept pursued his spiritual objective through his natural functions and inclinations, which were sublimated and then gratified in rituals in order to disintegrate his normal personality. This implies that cultic life was also largely interiorized and that the whole world, because it became completely ritualized, was given a new and esoteric meaning.

Tantric worship (puja) is complicated and in many respects different from the conventional ceremonies that it has often influenced. Tantric devotees distinguish between an "external" and an esoteric meaning of their texts and interpret their texts by means of an ambiguous "twilight" language. Tantrists describe states of consciousness with erotic terminology and describe physiological processes with cosmological terminology. They proceed from "external" to "internal" worship and adore the Goddess mentally, offering their hearts as her throne and their self-renunciation as "flowers."

According to Tantrism, concentration is intended to evoke an internal image of the deity and to resuscitate the powers inherent in it so that the symbol changes into mental experience. This "symbolic ambiguity" is also much in evidence in the esoteric interpretation of ritual acts performed in connection with images, flowers, and other cult objects and is intended to bring about a transfiguration in the mind of the adept.

The mantras (sacred utterances, such as *hūṃ, hrīṃ,* and *klaṃ*) are an indispensable means of entering into contact with the power they bear and of transcending normal mundane existence. Most potent are the monosyllabic, fundamental, so-called *bīja* ("seed") mantras, which constitute the main element of longer formulas and embody the essence of divine power as the eternal, indestructible prototypes from which anything phenomenal derives its existence. The cosmos itself owes its very structure and harmony to them. Also important is the introduction of spiritual qualities or divine power into the body (*nyāsa*) by placing a finger on the relevant spot (accompanied by a mantra).

Those Tantrists who follow the "right-hand path" attach much value to the yoga that developed under their influence and to bhakti and aspire to union with the Supreme by emotional-dynamic means, their yoga being a self-abnegation in order to reach a state of ecstatic blissfulness in which the passive soul is lifted up by divine grace.

There is also a Tantric mantra yoga (discipline through spells), which operates with formulas, and a hatha-yoga, (Sanskrit: "union of force"). In addition to normal yogic practices—abstinences, observances, bodily postures, breath control that requires intensive training, withdrawal of the mind from external objects, concentration, contemplation, and identification that are technically helped by mudras (*i.e.,* ritual intertwining of fingers or gestures expressing the metaphysical aspect of the ceremonies or the transformation effected by the mantras) and muscular contractions—hatha-yoga consists of internal purifications (*e.g.,* washing out stomach and bowels), shaking the abdomen, and some forms of self-torture. The whole process is intended to "control the 'gross body' in order to free the 'subtle body.' "

Mudras

Some Tantrists also employ *laya yoga* ("reintegration by mergence"), in which the female nature-energy (representing the shakti), which is said to remain dormant and coiled in the form of a serpent (*kuṇḍalinī*) representing the uncreated, is awakened and made to rise through the six centres (chakras) of the body, which are located along the central artery of the subtle body, from the root centre to the lotus of a thousand petals at the top of the head, where it merges into the Puruṣa, the male Supreme Being. As soon as the union of shakti and Puruṣa has become permanent, according to this doctrine, wonderful visions and powers come to the adept, who then is emancipated. Some of the Tantric texts also pursue worldly objectives involving magic or medicine.

Tantric and Śākta views of nature, man, and the sacred. The Tantric movement is sometimes inextricably interwoven with Śāktism. Śāktism consists of doctrines and practices that assume the existence of one or more shaktis. These are "creative energies" that are inherent in and proceed from God and are also capable of being imagined as female deities. Shakti is the deciding factor in

Shakti

By courtesy of (centre) the Cleveland Museum of Art, Ohio, Mr. and Mrs. William H. Marlatt Fund; photographs, (left and right) Pramod Chandra



Various roles of Śakti, the female aspect of the divine.
(Left) Pārvatī, the beneficent mother, bronze statue, south Indian, *c.* 900. Formerly in the collection of Srinivasa Gopalachari. (Centre) Durgā, the destroyer, detail from a Basohli school painting, *c.* 1700. In the Cleveland Museum of Art, Ohio. (Right) Ardhanārīśvara, united with lord Śiva as half-male, half-female, sandstone sculpture from Jhālawār, 6th century. In the Jhālawār Archaeology Museum, Rājasthān, India.

the salvation of the individual and in the processes of the universe because God acts only through his energy, which, personified as a goddess, is his spouse. Her role is very different in the various systems: she may be considered the central figure in a philosophically established doctrine, the dynamic aspect of *brahman,* producing the universe through her maya, or mysterious power of illusion; a capricious demoniac ruler of nature in its destructive aspects; a benign mother goddess; or the queen of a celestial court. There is a comprehensive Śāktism that identifies the goddess (usually Durgā) with *brahman* and worships her as the ruler of the universe by virtue of whom even Śiva exists. As Mahāyoginī ("Great Mistress of Yoga"), she produces, maintains, and reabsorbs the world. As the eternal Mother, she is exalted in the *Devīmāhātmya* ("Glorification of the Goddess") section of the *Mārkaṇḍeya-Purāṇa* (an important medieval Śākta encyclopaedic text). In the Bengal cult of the goddess Kālī, she demands bloody sacrifices from her worshipers lest her creative potency fail her; this cult propounds the belief that birth and death are inseparable, that joy and grief spring from the same source, and that the frightening manifestations of the divine should be faced calmly.

The Vedic goddess Vāc (her name means "Word") was then already the energetic and productive partner of Prajāpati. As Ardhanārīśvara (the "Lord Who Is Half Female"), Śiva presides over procreation. The Śāktas—often markedly associated with Śaivism—drew the following conclusions: creation is the result of the eternal lust of the divine couple; the man who is blissfully embraced by a beloved woman who is Pārvatī's counterpart assumes Śiva's wonderful personality and, liberated, will continue the joy of amorous sport.

Vaishnava Śāktism

In all of his incarnations Vishnu is united with his consort, Lakṣmi. The sacred tales of his various relations with her manifestations led his worshipers to view human devotion as parallel to the divine love and hence as universal, eternal, and sanctified. In Vaishnava Tantrism, Lakṣmī plays an important part as God's shakti—that is, as a central metaphysical principle. In his supreme state, Vishnu and his shakti are indissolubly associated with one another so as to constitute the personal manifestation of the supreme *brahman,* also called Lakṣmī-Nārāyaṇa. In mythical imagery, Lakṣmī never leaves Vishnu's bosom. In the first stage of creation, she awakens in her dual aspect of action-and-becoming, in which she is the instrumental and material cause of the universe; Vishnu himself is the efficient cause. In the second stage, her "becoming" aspect is manifested in the grosser forms of the souls and the power of maya, which is the immaterial source of the universe. In displaying her power she takes into consideration the accumulated karma of the beings, judging mundane existence as merit and demerit. Presented in myth as God's wife and the queen of the universe, she is always intent on liberating, by her favour and compassion, the incarnated souls of the devout; that is, she allows them to reenter into herself because they are really "parts," or rather "contractions," of her own essence. After entering her, the liberated soul takes part in the perfect embrace of the divine couple. Pāñcarātra Vaishnavism emphasizes that Lakṣmī—who in the mythological sphere intercedes with her husband for the preservation of the world—spontaneously and by virtue of her own power differentiates herself from Vishnu because she has in view the liberation of the souls. This current of thought complicated its explanation of the relation between God and the universe—which was at the same time an attempt at assigning to God's manifestations a place in a harmonious theological and cosmological system—with an evolutionist theory of successive creations. God is assumed to manifest himself also in three other figures, mythologically his brothers, who, each with his own responsibility, have not only a creative but also an ethical function, by which they assist those who seek to attain to final emancipation.

Tantric ritual and magical practices. The ritual of the left-hand Tantrists consisted of a kind of black mass in which all of the taboos of conventional Hinduism were conscientiously violated. Thus, in place of the traditional five elements (*tattvas*) of the Hindu cosmos, these Tantrists used the five "m"s: *māṃsa* (flesh, meat), *matsya* (fish), *madya* (fermented grapes, wine), *mudrā* (frumentum, cereal, parched grain, or gestures), and *maithuna* (fornication). This latter element was made particularly antinomian through the involvement of forbidden women, such as one's sister, mother, the wife of another man, or a low-caste woman, who was identified with the Goddess. Menstrual blood, strictly taboo in conventional Hinduism, was also used at times. Such rituals, which are described in Tantric texts and in tracts against Tantrists, made the Tantrists notorious. It is likely, however, that the rituals were not regularly performed except by a relatively small group of highly trained adepts; the usual Tantric ceremony was purely symbolic and even more fastidious than the pujas in Hindu temples.

The cult of the Śāktas is based on the principle of the ritual sublimation of natural impulses to maintain and reproduce life. Śākta adepts are trained to direct all their energies toward the conquest of the Eternal. The ritual satisfaction of lust and the consumption of consecrated meat or liquor are esoterically significant means of realizing the unity of flesh and spirit, of the human and the divine. They are not considered sinful acts but, on the contrary, effective means of salvation. Ritual copulation—which may also be accomplished symbolically—is, for both partners, a form of sacralization, the act being a participation in cosmic and divine processes. The experience of transcending space and time, of surpassing the phenomenal duality of spirit and matter, of recovering the primeval unity, the realization of the identity of God and his Śakti, and of the manifested and unmanifested aspects of the All, constitute the very mystery of Śāktism. The interpretation—metaphorical or literal—of the doctrines is, however, largely a matter of opinion and practice. Ritual practice is indeed as varied as the doctrines. Extreme Śākta communities perform the secret nocturnal rites of the *śrīcakra* ("wheel of radiance," described in the *Kulārṇava Tantra*), in which they avail themselves of the natural and esoteric symbolic properties of colours, sounds, and perfumes to intensify their sexual experiences. Most Tantrists, however, eliminate all but the verbal ritual.

Individual and collective yoga and worship, conducted daily, fortnightly, and monthly "for the delectation of the deity," are of special importance. After elaborate purifications, the worshipers—who must be initiated, full of devotion toward the guru and God, have control over themselves, be well prepared and pure of heart, know the mysteries of the scriptures, and look forward to the adoration with eagerness—make the prescribed offerings, worship the mighty puissance of the Divine Mother, and recite the relevant mantras. Once they have become aware of their own state of divinity, they are qualified to unite sexually with the Goddess. If a woman is, in certain rituals, made the object of sexual worship, the Goddess is first invoked into her; the worshiper is not to cohabit with her until his mind is free from impurity and he has risen to divine status. Copulation with a low-caste woman helps to transcend all opposites; union with a woman who belongs to another man is often preferred because it is harder to obtain, nothing is certain in it, and the longing stemming from the separation of lover and beloved is more intense—it is pure *preman* (agape, or divine love); adoration of a girl of 16 aims at securing the completeness and perfection of which this number is said to be the expression. However, the texts reiterate how dangerous these rites are for those who are not initiated; those who perform such ritual acts without merging their minds in the Supreme are likely to go to one of the hells.

Bengali Vaishnava-Sahajiyā cult

The esoteric Vaishnava-Sahajiyā cult, which arose in Bengal in the 16th century, was another emotional attempt at reconciling the spirit and the flesh. Displaying contempt for social opinion, its adherents, using the natural (*sahaja*) qualities of the senses and stressing the sexual symbolism of Bengal Vaishnavism, reinterpreted the Rādhā–Krishna legend and sought for the perpetual experience of divine joy: because Krishna's nature is love and the giving of love and because man is identical with Krishna, the realization of love can, after an arduous training, be experienced in man's nature. Women, being a ritual necessity as well

as the embodiment of a theological principle, could even become spiritual guides, like Rādhā, conducting the worshiper in his search for realization. After reaching this state, he remains in eternal bliss, can dispense with guru and ritual, and be completely indifferent to the world, "steadfast amidst the dance of maya."

Tantric and Śākta ethical and social doctrines. These ethical and social principles, though fundamentally the same as those promulgated in the classical dharma works, breathe a spirit of liberality: much value is set upon family life and respect for women (the image of the Goddess); no ban is placed on traveling (conventionally regarded as bringing about ritual pollution) or on the remarriage of widows. Although Tantric and Śākta traditions did not oblige their followers to deviate in a socially visible way from the established order, they provided a ritual and a way of life for those who, because of sex or caste, could not participate satisfyingly in the conventional rites.

The ancient Tantric tradition, based on the esoteric tantra literature, has become, through time, so interwoven with more orthodox Hinduism that it is difficult to define precisely. Although it sees an identity between the soul and the cosmos, it speaks of the internalization of the cosmos rather than of the release of the soul to its natural state of unity. The body is the microcosm, and the ultimate state is not only omniscience but total realization of all universal and eternal forces. The body is real, not because it is the function or creation of a real deity but because it contains the deity, together with the rest of the universe. The individual soul does not unite with the One—it is the One, and the body is its function.

Tantrism, though not always in its full esoteric form, is a feature of much modern mystical thought. In Tantrism the consciousness is spoken of as moving—driven by repetition of the mantra and by other disciplines—from gross awareness of the material world to realization of the ultimate unity. The image is of a serpent, coiled and dormant, awakened and driven upward in the body through various stages of enlightenment until it reaches the brain, the highest awareness. The modern mystic Ramakrishna describes the process, which also describes the experience that all Hindu mystical processes seek:

The coiled serpent

> When [the serpent] is awakened, it passes gradually through [various stages], and comes to rest in the heart. Then the mind moves away from [the gross physical senses]; there is perception, and a great brilliance is seen. The worshiper, when he sees this brilliance, is struck with wonder. The [serpent] moves thus through six stages, and coming to [the highest one], is united with it. Then there is *samādhi*. . . When [the serpent] rises to the sixth stage, the form of God is seen. But a slight veil remains; it is as if one sees a light within a lantern, and thinks that the light itself can be touched, but the glass intervenes. . . . In *samādhi*, nothing external remains. One cannot even take care of his body any more; if milk is put into his mouth, he cannot swallow. If he remains for twenty-one days in this condition, he is dead. The ship puts out to sea, and returns no more. (Translation by Edward C. Dimock, Jr. Source is *Śrīśrīrāmakṛṣṇa-kathāmṛta;* Calcutta: Ramakrsna Mission.)

VERNACULAR LITERATURES

Most of the texts cited in this survey are Sanskrit texts, which constitute the oldest layer of preserved Hindu literature. But the sacred literature of India is by no means as monolithic as these texts might suggest. Several other essential elements exist: independent sacred literatures in languages other than Sanskrit and material in other languages related to the Sanskrit texts either as sources of material now preserved only in Sanskrit or as new texts originating as translations of Sanskrit texts. Because Sanskrit has been in intimate contact with the "mother tongues" of India for such a long time, it is often impossible to determine in which of these categories a particular vernacular text belongs.

Indologists usually emphasize the influence of Sanskritic (often called "Aryan") culture on Dravidian culture, and indeed this influence was considerable. Sanskritic influence was already in evidence in the earliest Tamil (a principal Dravidian language) literature, perhaps dating from the 1st century AD. At this time in South India the orthodox cults were aristocratic in character and were supported by kings and chiefs who gained in prestige by patronizing Brahmans and adopting Aryan ways. The Tamils were still primarily devoted to the old cults, some of which, however, were taking on an Aryan complexion. The pastoral god Murugan was identified with Skanda and his mother, the fierce war goddess Koṟṟavai, with Durgā. Varuṇan, a sea god who had adopted the name of the old Vedic god but otherwise had few Aryan features, and Māyon, a black god who was a rural divinity with many of the characteristics of Krishna in his pastoral aspect, also are depicted in Tamil literature. The final Sanskritization of the Tamils was brought about through the patronage of the Pallava kings of Kānchipuram, who began to rule in the 4th century AD and who financed the making of many temples and fine religious sculptures. Similar processes were taking place in the Deccan, Bengal, and other regions.

Sanskritization is a term that refers to a style of text that imitates the customs and manners of the Brahmans. But, although most sacred texts in Sanskrit were composed by Brahmans, many were also composed by lower-class authors. Likewise, although some sacred texts in vernacular languages were written by authors of lower castes, many others were written by Brahmans. In addition, because Sanskrit ceased to be spoken as a primary language soon after the Vedas were composed, it is likely that most of the thoughts underlying all subsequent Sanskrit literature were first thought in some other language. Yet Indologists tend to be Sanskritists, and Sanskritists tend to assume that all texts originated in Sanskrit. Indeed, even the counterbalancing tendency to acknowledge the flow from non-Sanskrit to Sanskrit sources has often misfired; far too often it is merely asserted that anything that appears in post-Vedic Hinduism and is not attested in the Vedas is "Dravidian," or, even worse, from the Indus civilization (about whose religion virtually nothing is known).

The issue is further clouded by the fact that, though Sanskrit texts tend to be written and vernacular traditions are primarily oral, there are important oral traditions in Sanskrit, too (including the traditions of the two great Sanskrit epics), and there are important manuscript traditions in some of the non-Sanskritic languages (such as Bengali and Tamil). Indeed, written and oral versions of the epics and *Purāṇas* have been, from the very start, in constant symbiosis.

Little relevance, therefore, attaches to a distinction between "written" and "oral" traditions. A myth is essentially told or narrated, a process that is designated in Sanskrit by such words as *purāṇa* (ancient story) and *ākhyāna* (illustrative narrative). In the oldest source, the Rigveda, myths are not so much told as alluded to; it is in the later Vedic literature of the Brāhmaṇas that narratives are found, and these are often prejudiced by liturgical concerns of the authors.

Oral and written Hindu myths

The recitation of certain myths was prescribed for specific rituals. The epic *Mahābhārata* states that Vedic stories were narrated "in the pauses of the ritual," probably by Brahmans. The warrior class (Kṣatriyas) had their own mythographers in their *sūtas* (charioteers and panegyrists), who celebrated the feats of great rulers. These *sūtas*, who became popular narrators of myth and legend, had their own bardic repertoire, which soon was extended to higher mythology. They—and other wanderers who found ready audiences at sacrifices or places of pilgrimage—disseminated the lore.

Such narrators still continue to repeat and embroider their ancient stories of gods, sages, and kings. At an early stage their narratives were dramatized and gave rise to the Sanskrit theatre, in which epic mythic themes preponderate, and to the closely related dance, which survives in the now largely South Indian schools of bharata natya (traditional dance) and the kathakali (narrative dance) of Kerala. Thus, even in Sanskrit literature, oral performance was an essential component, which further facilitated the assimilation of oral vernacular elements.

When the Indo-Europeans, who spoke Sanskrit, an Indo-European language, entered India in around 1500 BC, most of the people they encountered spoke languages that belonged to a major non-Indo-European linguistic group

called Dravidian. These two language groups interacted from a very early period, and although the earliest preserved specimens of Sanskrit (themselves dating from a period long after the Indo-Aryan invasions) far antedate examples of any other languages, there is good reason to believe that the other languages also produced texts, although unrecorded, at a very early period. When the devotional aspect of Hinduism came into full flower, the vernacular traditions both in Dravidian language groups and in languages derived from Sanskrit began to record their texts and to have a more discernible influence upon the Sanskrit tradition.

Of the four primary Dravidian literatures—Tamil, Telugu, Kannada, and Malayalam—the oldest and best-known is Tamil. The earliest preserved Tamil literature, the so-called *Caṅkam* or *Śaṅgam* poetry anthologies, dates from the 1st century BC. These poems are classified by theme into *akam* ("interior," primarily love poetry) and *puṟam* ("exterior," primarily about war, the poverty of poets, and the deaths of kings). The bhakti movement has been traced to Tamil poetry, beginning with the poems of the devotees of Śiva called Nāyaṇārs and the devotees of Vishnu called Āḻvārs. The Nāyaṇārs, who date from about AD 800, composed intensely personal and devout hymns addressed to the local manifestations of Śiva.

The most famous Nāyaṇār lyricists are Appar, Sambandar, and Cuntarar, whose hymns are collected in the *Tevāram* (*c.* 11th century). More or less contemporary were their Vaishnava counterparts, the Āḻvārs Poykai, Pūtaṉ, Pēyār, and Tirumaṅkaiyāḻ-vār, and in the 8th century the poetess Āṇḍaḷ, Periyāḻvār, Kulacēkarar, Tiruppāṇāḻvār, and notably Nammāḻvār, who is held to be the greatest. The devotion of which they sing exemplifies the new bhakti movement that seeks a more direct contact between man and God, carried by a passionate love for the deity, who reciprocates by extending his grace to man. These saints also became the inspiration of theistic systematic religion: the Śaivas for the Śaivā-siddhānta, the Vaishnavas for Viśiṣṭādvaita. In Kannada the same movement was exemplified by Basava, whose *vācanams* ("sayings" or "talks") achieved great popularity. His religion, that of Virašaivism, was perhaps the most "protestant" of the bhakti religions.

Dravidian traditions

New Dravidian genres continued to evolve into the 17th and 18th centuries, when the Tamil Cittars (from the Sanskrit *siddhas*, "perfected ones"), who were eclectic mystics, composed poems noted for the power of their naturalistic diction. The Tamil sense and style of these poems belied the Sanskrit-derived title of their authors, a phenomenon that could stand as a symbol of the complex relationship between Dravidian and Sanskrit religious texts.

The main languages derived from Sanskrit are Bengali, Hindi (with its many dialects, of which Maithili is the oldest and Urdu, heavily influenced by Persian and Arabic and written in a Perso-Arabic script, is the most important), Punjabi, Gujarati, Marāṭhi, Oriya, Kashmiri, Sindhi, Assamese, Nepali, Rajasthani, and Sinhalese. Most of these languages began to develop literary traditions around AD 1000. The earliest texts in Hindi are those attributed to the 13th–14th-century Muslim poet Amīr Khosrow.

Hindi literature produced its own great religious lyricists beginning with the disciples of Rāmānanda (*c.* 1450), who was a follower of the philosopher Rāmānuja. Among them the most famous is Kabīr, whose bhakti was nonsectarian. Tulsīdās, apart from his *Rāmcaritmānas*, composed Rāmaite lyrics. Sūrdās (1483–1563), a follower of the Vallabha school of Vedānta, is famous for his *Sūrsāgar* ("Ocean of the Poems of Sūr"), a collection of poems based on the childhood of Krishna, following the account of the *Bhāgavata-Purāṇa*. In the Marāṭhi tradition, Nāmdev (*c.* 1300) celebrated Vishnu, particularly in his manifestation as Viṭṭhobā at the Pandharpur temple; and in the 17th century Tukārām, the greatest poet of this literature, sang of the god of love in numerous hymns.

A small sect, the Kabīrpanthis, acknowledges Kabīr as its founder, but its importance is less than that of the vigorous new religion (Sikhism) founded by one of Kabīr's disciples, Nānak. In its final form, Sikhism contains elements taken from Islām (equality in the faith, opposition to iconolatry, extreme reverence for the sacred book) and probably also from Christianity (the Sikh baptism and communion meal), but its theology is still essentially Hindu.

Although the earliest Hindu text in Bengali is a mid-15th-century poem about Rādhā and Krishna, medieval texts in praise of gods and goddesses, known as *maṅgalkāvyas*, must have existed in oral versions long before that. In later Bengal Vaishnavism, the emphasis shifts from service and surrender to mutual attachment and attraction between God (*i.e.*, Krishna) and humankind: God is said to yearn for the worshiper's identification with himself, which is his gift to the wholly purified devotee. The mystical and devotional possibilities of the Krishna legend are made subservient to religious practice; the divine sport and wonderful feats of this youthful hero are interpreted symbolically and allegorically. Thus, the highest fruition of bhakti is admission to the eternal sport of Krishna and his beloved Rādhā, whose sacred love story is explained as the mutual love between God and the human soul. Various gradations of bhakti are distinguished, such as awe, subservience, and parental affection. These are correlated with the persons of the Krishna legend; the highest and most intimate emotion is said to be the love of Rādhā and her girlfriends for Krishna.

A particularly rich tradition centred in Bengal concentrated on the love of Rādhā, who symbolizes the human soul, for Krishna, the supreme God. In this tradition are Caṇḍidās and the Maithili poet Vidyāpati (*c.* 1400). The greatest single influence was Caitanya, who in the 16th century renewed Krishnaism. He left no writings but inspired many hagiographies, among the more important of which is the *Caitanya-caritāmṛta* ("Nectar of Caitanya's Life") by Krishna Dās (born 1517).

The work of Caitanya

Caitanya had a profound and continuing effect on the religious sentiments of his Bengali countrymen and propagated the community celebration (*saṃkīrtana*) of Krishna as the most powerful means of bringing about the proper bhakti attitude. Caitanya also introduced the worship of God, the director of man's senses, through the very activity of man's senses, which must be free from all egoism and completely filled with the intense desire (*preman*) for the satisfaction of the beloved (*i.e.*, Krishna).

The religious lyric continues in the so-called *pādas* (verses); one of the greatest poets in this bhakti genre in which divine love is symbolized by human love is Govinda Dās (1537–1612). The songs of Rāmprasād Sen (1718–75) similarly honour Śakti as mother of the universe and are still in wide devotional use. The most famous religious lyrics in Gujarati are the poems of the saint Mīrā Bāī (1503–73), who wrote passionate love poems to Krishna, whom she regarded as her husband and lover.

The complex interaction between Sanskrit and non-Sanskrit religious classics may be seen in the development of the epics. Parts of the two great Sanskrit epics, the *Mahābhārata* and the *Rāmāyaṇa*, and many *Purāṇas* (especially the *Bhāgavata-Purāṇa*) were translated into various vernaculars. Technically, these works were not literal translations, but free versions in which the authors placed their own emphases, different from the original and from one another. The oldest of the vernacular versions of the *Rāmāyaṇa* is the Tamil one of Kampaṉ (*c.* 12th century), a work suffused with devotion (bhakti) and of high literary distinction. Another famous translation in Tamil, written by Villiputturar, exists from the 18th century. A Telugu rendering was made by Ranganātha about 1300. In Bengali several translations were made, with some interesting and probably authentic variations from the "official" Rāma story by Vālmīki, the best-known one by Kṛttibās Ojhā (1450). Equally, if not more, famous is the Hindi version by Tulsīdās (*c.* 1550), entitled *Rāmcaritmānas* ("Holy Lake of the Acts of Rāma").

Regional language versions of the *Rāmāyaṇa* and *Mahābhārata*

The *Mahābhārata* was translated into Bengali about 1600 and into Telugu by Nannaya and Tikkana in the 13th century. The *Bhāgavata-Purāṇa*, which was translated frequently (*e.g.*, into Bengali by Maladhar Vasu, 1480), was popular both as a text and because it gave the canonical account of Krishna's life and especially his boyhood, which is the perennial inspiration of the bhakti poets.

In Marāṭhi the teacher Jñānadeva (also known as

Jñāneśvara; *c.* 1275–96) composed a commentary on the *Bhagavadgītā* that remains a classic in that literature. His work was continued by Eknāth (*c.* 1600), who also composed bhakti poetry. In the 16th century the Kannada poet Gadugu produced his own highly individual version of the *Mahābhārata.* In addition to the above literal or not-so-literal translations of the Sanskrit epics, the Tamils composed their own epics, notably Iḷaṅkō Aṭikaḷ's *Cilappatikāram* ("The Lay of the Anklet") and its sequel, the *Maṇimekhalai* ("Jeweled Girdle"). In Telugu there is the great Palnāḍu Epic; Rajasthani has an entire epic cycle about the hero Pabuji. The remaining vernaculars have produced many other works of the epic genre.

Myth and modernity

Hindu mythology in contemporary India. Much of the classical mythology persists today, and Hindus are exposed to it year-round. Meanwhile, the mass media have made their contributions: the type of motion picture called "mythological" is extremely popular, perpetuating the ancient stories to the village level, and so are "devotionals," in which an example of bhakti is illustrated. The radio regularly carries *bhajans* (devotional songs) and classical South Indian songs, the themes of which are often mythic. Every orthodox Hindu's home has at least one corner set aside as a domestic sanctuary where representations of a chosen deity are placed, and puja (worship) is done with prayers, hymns, flowers, and incense. Richer establishments set aside entire rooms as shrines. Mythic illustrations are favourites in Indian calendar art.

Mythology has adjusted itself effortlessly to modernity. The ashram of the 20th-century mystic and religious leader Śrī Aurobindo in Pondicherry, dedicated to the Mother Goddess (personified by this group as a single principle), is an extremely modern establishment complete with tennis courts. New temples are constructed with modern techniques; one temple in Vārānasi contains mirrors onto which are etched the entire *Rāmcaritmānas.* This same poem is the basis of the annual celebration of Rām Lilā (the play of Rāma) in northern India, in which the entire community participates. The Rāma story was evoked by Mahatma Gandhi when he set the Rām Rāj ("Kingdom of Rāma") as India's governmental ideal. On occasion, social protesters arm themselves with myth to make a point. For example, the personality of Karṇa, an antagonist in the *Mahābhārata* who is berated for his low birth, is extolled in intellectual circles as a truer champion than the aristocratic heroes. A Kannada-language play of the 1960s based on the life of King Yayāti enjoyed great popular and critical success. Anti-northern groups in Tamil Nadu revised the story of Rāma, whose expedition against the demon Rāvaṇa is believed by some to be the Aryan invasion of South India, by reversing it to abuse Rāma and to glorify Rāvaṇa.

On a popular level, people at temples and fairs are continually reacquainted with their mythological heritage by *paurāṇikas,* tellers of the ancient stories, heirs of the *sūtas* of 3,000 years ago, and no festival ground is complete without tents where the religious are reminded of their myths by pious speakers, modestly compensated by fees but richly rewarded by the honour in which they are held.

FOLK HINDUISM

Despite the impact of the West, the propaganda of modern reform movements, and the spread of education and secularist modernization, Hinduism has changed only slowly. For ordinary Hindus, religion primarily consists of the manual and verbal performance of rites to promote their private interests. The innumerable ceremonies, observances, fasts, feasts, pilgrimages, and visits to nearby temples constitute the essence of religion.

General characteristics of folk traditions. For millions, the main motive of religious practices is still the fear of ambivalent powerful beings. Most Hindus propitiate the meat-eating, sometimes benevolent but largely malevolent deities concerned with man's daily events, their ancestors or the founder of their community, and those various spirits that have no permanent residence and cause evil and misfortune. Hindus strive to escape the powers of the evil eye; to manipulate those spirits dwelling in wells, trees, stones, water, and ground; to counteract curses, witchcraft, plague, and cholera; and to worship village godlings who may give rain or a bountiful harvest. They make use of astrology, divination, and the reading of omens and auspicious moments. A large variety of purifications and ritual prohibitions, charms, and amulets to ward off any kind of misfortune (including bad luck in lawsuits and examinations) are, in the eyes of the majority, of greater importance than the atman-*brahman* doctrine. Even the hope of heaven or the fear of hell has little vogue in various regions, except among the higher castes.

It is difficult to draw a sharp line of distinction between popular Hinduism, the beliefs and practices of more or less Hinduized "external" groups, and Indian tribal religion. Many elements of tribal culture that in a particular region have not been adopted by those recognized as belonging to the Hindu fold are in fact similar to what has been adopted by Hindus in other areas. Tribal people and outcaste groups are, on the other hand, always willing to worship a few more gods or to imitate the rituals of lower-caste Hindus. Age-long processes of interpenetration and fusion have led to an adoption of many local and popular cults into general Hinduism—or, because it expressed itself mainly in Sanskrit, into the Great, or Sanskritic, tradition—and to the identification of regional gods with the great figures of the Hindu pantheon. Popular belief is integrated rather than discounted or discarded. This process is facilitated by a tendency toward the assimilation of local beliefs by pan-Indian Hinduism and by an unwillingness to deny gods and cults (the worship of a local river deity, for example, may be identified with that of the Ganges). The inheritors of the Little, or regional, traditions accepted, as a result of continual and complicated Hinduizing influences, vegetarianism, regular fasts, and food restrictions; they also began to object to the remarriage of widows, to observe Hindu festivals, to sing Hindu religious songs, to perform funeral and other ceremonious worship and, most importantly, to imbibe the ideas embodied in religious and mythological narratives. Thus, various tribal or outcaste groups have a religion with some affinities to a simple Śaivism without sacred books or regular liturgy.

Relation of folk Hinduism to Indian tribal religions

While many Hindus pursue their approach to the divine individually, corporate worship in families, villages, and sects is far more common in some castes. These groups exhibit the utmost variation in beliefs and practices. As a rule, each community practices only a small segment of the whole spectrum of religious behaviour as the expression of its own religious life. As to the relation between religion and social structure, there are in many communities differently structured systems, each with its own religious behaviour, in which their members may be involved. As members of a joint family, they take part in the domestic cult and ritually express family solidarity at such critical points as mourning or marriage; as members of a village community, they take part in its particular cult, which is a collective action of that community. Different castes, however, establish their own rituals in order to foster unity and to differentiate themselves from others. There is, on the other hand, ritual cooperation between different villages of the same region.

For many communities, spiritual reality is complex: while many women may address local spirits, family ancestors, and goddesses of disease, some of the men may embrace monotheistic ideas. The village's guardian spirit and the saint of a Muslim shrine may also be worshiped, Rāma's name is invoked in prayers, and the major deities are honoured chiefly during their periodic festivals. Marriage and other ceremonies combine ancient Sanskritic rites with popular and local features, and even members of the higher classes may accept the entire range of belief. In many regions, each caste has both general Hindu and "parochial" rituals and beliefs, but the proportions in which the two are found together vary from caste to caste and from locality to locality. The upper castes everywhere, however, have a certain amount of Sanskritic ritual in common; but even those who are more or less exclusively devoted to Śiva, for example, do not necessarily constitute a Śaiva community.

The bhakti movements have long influenced the religious feelings of their followers, and religious problems

are discussed by people of all professions and intellectual levels. Divine assistance is implored on every imaginable occasion; ancient Vedic rites have even been used as a defense against atomic danger.

Regional varieties of folk religion. In the hilly and mountainous regions of North India, Śaivism, aligned with Śāktism, is prevalent. The awe and mystery of the jungle and mountains are, there and elsewhere, personified as forest "Mothers" or mountain deities, represented by piles of stones or branches of trees to which every passerby contributes an offering. Mother Earth is a great goddess whose marriage (with the Earth god or the Sun) is festively celebrated and whose annual period of impurity is observed by a cessation of all agricultural activities. During the harvest season she is propitiated with wild orgies. In secluded parts of central India she is identified with Devī, a goddess mostly worshiped in North India. Very often, however, she shows herself in her malevolent form, as Mother Death (Mārī) or as Kālī. There are also many lower caste groups that have adopted a Vaishnava way of life either in order to raise their social status or to have a prospect of salvation.

Folk goddesses

In the east and northeast, where, broadly speaking, Śāktism is dominant, though Vaishnavism is also common, popular belief has modified the transmigration doctrine by the assumption that the soul of the deceased reappears in a child born in the same family within a year after the person's death. Among the female deities, there are tutelary goddesses of young children and women in childbed: Ṣaṣṭhī, "the Sixth," is worshiped on the sixth day after birth and is represented by a compost pile of cow dung or earth that is placed in the birth room; and Caṇḍī, a form of the goddess Durgā, lives in trees and is propitiated by lumps of earth. The snake goddess Manasā is personified in a plant of the same name or in a stone carved into the shape of a female seated on a snake; a day in the rainy season, when reptiles are most dangerous, is devoted to her priestless and unpretentious worship. In literary works she is eulogized as the Great Mother who is expected to give a prosperous journey through life and, to a certain extent, is also Sanskritized by being identified with epic snake demons. Another example of fusion of general and local Hindu institutions is the conviction that ghosts and demons are warded off by performing a ceremony in honour of the deceased at Gayā in modern Bihār state.

In many regions—especially in western India, where Vaishnavism is dominant—believers admit that virtue will improve their lot in a subsequent existence, but they do not seem to strive for final union with the Supreme. Here, and elsewhere, a workaday religion meant to meet the requirements of everyday life exists alongside a higher religion understood only by the Brahmans, who are called on to officiate on important occasions. In order to discover the divine will, exorcists and mediums, possessed by mother goddesses and submitting themselves to self-torture, are called upon to prophesy about future events. In these regions the worship of snakes is more prominent, and some temples are even dedicated to them. Practices based on the belief in scapegoats, ritual nudity, and black magic are also widespread.

The whole of peninsular India is mainly devoted to Śaivism, devotional forms of Vaishnavism, and the worship of the goddess in her many forms. A striking feature in the religion of South India is the propitiation of usually local female village deities of varied and ambivalent character, to whom in almost every settlement a simple shrine or other sacred place is dedicated. These deities are thought to be particularly competent for dealing with the facts of village life, such as diseases of the inhabitants and their cattle. Special cholera and smallpox goddesses are the subject of elaborate stories. In a few cases—*e.g.*, that of Māriyammā, the smallpox goddess of South India—such a goddess is known to a large region. These mothers, from whom all good and bad luck emanate, are almost universally worshiped with animal sacrifices, and the priestly minstrants (*pūjārī*) officiating in their cult belong to the non-Brahman groups. The goddesses may be represented by various symbols (stone pillars, sticks, clay figures) that need not be permanent. Most of their shrines are simple, small brick buildings or rough stone platforms under a tree. Offerings of rice, fruit, and flowers may be made every day or on fixed days; although there often is a fixed annual festival, it is not uniformly celebrated and no calendar of festivals is established. An exception to this is the male deity Aiyanār, who in the countryside of Tamil Nādu state is worshiped as the watchman and patron of the villages but also is implored to grant children and other blessings. He is a vegetarian and therefore ranks as socially superior to the female village goddess with whom he has entered into a complementary relation. Aiyanār is worshiped either as a village deity or in a temple dedicated to Śiva, where he is given the rank of a son of that god. In these Śiva temples he is legitimated by higher Hinduism and fulfills the function of a double of Śiva representing "All-India" or general Hinduism in the village, which does not regard him as an outsider. Śiva himself is also worshiped and given a consort, who, though considered a manifestation of Durgā, has various names according to the tradition of temple or village.

Aiyanār

Sacred snakes, especially cobras, are also given offerings—partly to avert danger from these reptiles, partly to propitiate them with the aim of obtaining rain, fertility, or children; to that end women worship snake stones (*nāgal-kals*) or erect stone figures of cobras. Every joint family of the Coorgs in Karnātaka state and most other peoples have a snake deity of their own that is said to embody their welfare. Here and there, Brahmans officiate in this cult, which usually takes place in small sanctuaries in private gardens. Although also known in other parts of India, the methods of exorcising evil spirits known as devil-dancing are most fully developed in South India. The notorious hook-swinging festival, Caḍak-pūjā, held for propitiatory purposes in cases of famine or other calamities—a man was suspended by hooks at the end of a long pole and swung around—though strictly prohibited, survived to the 20th century. Greater festivals are, generally speaking, either celebrated at the chief agricultural seasons or connected with the expulsion of malign powers.

Folk and tribal myths. There is a great diversity in folk mythology throughout the entire Indo-Pakistan subcontinent, but these myths have neither been fully collected nor systematically studied. Among locally important deities, Manasā, a snake goddess, worshiped in Assam and Bengal to ward off snakebites and secure prosperity, has an enormous mythology of her own. In South India one finds popular cobra cults with a variety of myths and lore. In Mahārāshtra, a form of Vishnu, known as Viṭṭhal or Viṭṭhobā, has also spawned a rich mythology.

The sources of folk and tribal mythology are vernacular literature, oral tradition, folklore, and folk and tribal arts. Folk mythology derives from the most ancient times and has influenced both Vedic and classical mythology. Classical mythology became what it is by continuously assimilating myths not previously known or accepted, so that the line between classical and folk (and tribal) mythology is apt to be arbitrary. The Great (Sanskrit) tradition of classical mythology (as contrasted with the local, or Little, traditions of folk mythology) may include within its scheme a god who continues to have an independent existence on a folk level. Sacred manifestations of purely local interest are associated with the higher mythology by becoming a Little manifestation of a Great god, such as the footstep of Rāma and the bathing place of Sītā (Rāma's wife). Conversely, an incident of the Great tradition may be adopted and adapted on the folk level. For example, the local Mahārāshtrian god Viṭṭhobā is identified with a manifestation of Vishnu and thus assured a place in the Great tradition; on the other hand, in North India, the widely celebrated festival of Navarātrī ("Nine Nights") is associated with the village goddess Naurthā.

The concepts of avatar and *vāhana*

Certain concepts that evolved in Purāṇic mythology have facilitated the absorption of folk elements; two of these should be singled out: avatar (*avatāra,* incarnation), and *vāhana* (vehicle).

The concept of avatar (literally "descent") issues from the belief that in times of trouble a god, notably Vishnu, incarnates himself as a man or hero to set matters right. Such a concept provides the opportunity for identifying a

local deity (like Viṭṭhobā, above) with an all-Indian god like Vishnu. The concept may also extend to the worship of very local hierophanies (manifestations of the sacred; *e.g.*, South Indian Vaishnavism accepts "icon-incarnation" [*arcāvatāra*], in which Vishnu "descends" in a local icon).

According to the concept of a *vāhana* (literally "mount"), every god has an entourage of his own, which includes a favourite riding animal; this facilitates many folk associations. Vishnu's mount is the bird Garuḍa, an old solar symbol; Śiva's is the bull Nandi, whose worship may go back to the ancient Harappan civilization. There are other mythological patterns, such as adoption in a family; thus the folk god Gaṇeśa, an elephant-headed god, is made the son of Śiva, as is Kumāra Kārttikeya, the war god, who arose from the South Indian war god Muruganˉ. Hanumān, the monkey god, becomes an all-Indian god as a helper of Rāma, who is an avatar of Vishnu.

Other spirits and godlings of folk provenance are not absorbed to the same degree and thus retain their folk character. Important are the snakes (nagas), to which great power is attributed; the yakshas, koboldlike keepers of wealth, whose king is Kubera; *vetāla*s, ghoulish pranksters who haunt corpses; and spirits of the restless dead (bhuts, pretas), who must be warded off. Though the existence of these spirits is fully recognized in the classical mythology, they operate primarily on the folk level.

It is therefore difficult to find folk myths and cults that are not in one form or another elevated into the Sanskrit tradition by a specific association with a major god. Only where there has been no appreciable cultural contact between Indian tribal people and Hindus (or "Hinduized" folk groups) can a significant distinction be drawn between classical Hindu mythology and "folk" mythology. The myths of the tribes of Chota Nāgpur (Bihār state), the Santāl (West Bengal and Bihār states), the Toda in the Nilgiri Hills (Tamil Nādu state), and others are examples of such mythology. Much religious material lies hidden in folktales.

(J.A.B.v.B./E.C.D./A.L.B./W.Do./B.K.S./Ed.)

Rituals, social practices, and institutions

SACRIFICE AND WORSHIP

Although the Vedic fire rituals were largely replaced in Purāṇic and modern Hinduism by image worship and other forms of devotionalism, many Hindu rites can still be traced back to Vedism. Certain royal sacrifices—such as the *rājasūya*, or consecration ritual, and the horse sacrifice (*aśvamedha*)—remained popular with Hindu kings until very recently. Other large-scale Vedic sacrifices (*śrauta*) have been regularly maintained from ancient times to the present by certain families and groups of Brahmans. By and large, however, the surviving rituals from the Vedic period tend to be most clearly observed at the level of the domestic (*gṛhya*) ritual.

Domestic rites. The Vedic householder was expected to maintain a domestic fire into which he made his offerings. Normally he did this himself, but in many cases he employed a Brahman officiant. In the course of time, the family priest was given a large part in these ceremonies, so that most Hindus have employed Brahmans for the administration of the "sacraments" (samskara). The samskaras include all important life-cycle events, from conception to cremation, and are the main constituents of the domestic ritual.

Rites of passage

The sacraments are transitional rites intended to make a person fit for a certain purpose or for the next stage in life, by removing taints (sins) or by generating fresh qualities. If the blemishes incurred in this or a previous life are not removed, the person is impure and will acquire no reward for any ritual acts. The sacraments, while sanctifying critical moments, are therefore deemed necessary for unfolding a person's latent capacities for development. Śūdras are allowed to perform some samskaras if they do not require the use of Vedic mantras.

Samskaras: rites of passage. In antiquity there was a great divergence of opinion about the number of rites of passage, but in later times 16 were regarded as the most important. The impregnation rite, consecrating the supposed time of conception, consists of a ritual meal of pounded rice (mixed "with various other things according to whether the married man desires a fair, brown, or dark son; a learned son; or a learned daughter"), an offering of rice boiled in milk, the sprinkling of the woman, and intercourse; all acts are also accompanied by mantras. In the third month of pregnancy the rite called *puṃsavana* (begetting of a son) follows. The birth is itself the subject of elaborate ceremonies, the main features of which are an oblation of ghee (clarified butter) cast into the fire; the introduction of a pellet of honey and ghee into the newborn child's mouth, which according to many authorities is an act intended to produce mental and bodily strength; the murmuring of mantras for the sake of a long life; and rites to counteract inauspicious influences. There is much divergence of opinion as to the time of the name-giving ceremony; in addition to the personal name, there is often another one that should be kept secret for fear of sinister designs against the child.

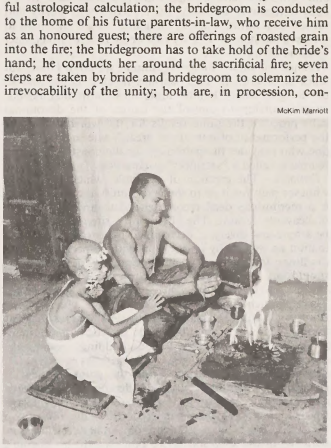
Rites associated with the newborn

In modern times most samskaras (with the exceptions of impregnation, initiation, and marriage) have in many areas fallen into disuse or are performed in an abridged or simplified form without Vedic mantras or a priest. This tendency was encouraged by the accommodating spirit of the Brahmans, who allowed their clients easy atonements for the nonobservance of rites. The important *upanayana* initiation is held when a boy is between the ages of eight and 12 and marks his entry into the community of the three higher classes of society. In this rite he becomes a "twice-born one," or *dvija*. Traditionally, this was also the beginning of a long period of Veda study and education in the house under the guidance of a teacher (guru). In modern practice, the haircutting ceremony—formerly performed in a boy's third year—and the initiation are usually performed on the same day, the homecoming ceremony at the end of the period of study being little more than a formality.

Marriage customs

Wedding ceremonies, the most important of all, have not only remained elaborate—and often very expensive—but have also incorporated various elements—among others, propitiations and expiations—that are not indicated in the oldest sources. Already in ancient times there existed great divergences in accordance with local customs or family or caste traditions. However, the following practices are usually considered essential. The date is fixed only after careful astrological calculation; the bridegroom is conducted to the home of his future parents-in-law, who receive him as an honoured guest; there are offerings of roasted grain into the fire; the bridegroom has to take hold of the bride's hand; he conducts her around the sacrificial fire; seven steps are taken by bride and bridegroom to solemnize the irrevocability of the unity; both are, in procession, con-

McKim Marriott



A young boy performs his first puja after initiation into the community of the "twice-born."

ducted to their new home, which the bride enters without touching the threshold.

Of eight forms of marriage recognized by the ancient authorities, two have remained in vogue: the simple gift of a girl and the legalization of the alliance by means of a marriage gift paid to the bride's family. In the Vedic period, girls do not seem to have married before they had reached maturity. Child marriage and the condemnation of the remarriage of widows, especially among the higher classes, became customary later and have gradually, since the mid-19th century, lost their stringency.

Death rites and customs

The traditional funeral method is cremation (a family affair), burial being reserved for those who have not been sufficiently purified by samskaras (*i.e.,* children) and those who no longer need the ritual fire to be conveyed to the hereafter, such as ascetics who have renounced all earthly concerns. An important and meritorious complement of the funeral offices is the sraddha ceremony, in which food is offered to Brahmans for the benefit of the deceased. Many people are still solicitous to perform this rite at least once a year even when they no longer engage in any of the five obligatory daily offerings.

Daily offerings. There are five obligatory offerings: (1) offerings to the gods (food taken from the meal); (2) a cursory offering (*bali*) made to "all beings"; (3) a libation of water mixed with sesame offered to the spirits of the deceased; (4) hospitality; and (5) recitation of the Veda. Although some traditions prescribe a definite ritual in which these five "sacrifices" are performed, in most cases the five daily offerings are merely a way of speaking about one's religious obligations in general.

Other private rites. The morning and evening adorations (*sandhyā*), being a very important duty of the traditional householder, are mainly Vedic in character, but they have, by the addition of Purāṇic and Tantric elements, become lengthy rituals. If not shortened, the morning ceremonies consist of self-purification, bathing, prayers, and recitation of mantras, especially the Gāyatrī mantra (Rigveda 3.62.10), a prayer for spiritual stimulation addressed to the Sun. The accompanying ritual includes (1) the application of marks on the forehead, characterizing the adherents of a particular religious community, (2) the presentation of offerings (water, flowers) to the Sun, and (3) meditative concentration. There are Śaiva and Vaishnava variants, and some elements are optional. The observance of the daily obligations, including the care of bodily purity and professional duties, leads to mundane reward and helps to preserve the state of sanctity required to enter into contact with the divine.

Gāyatrī mantra

Temple worship. Image worship in sectarian Hinduism takes place both in small shrines in each house and in the temple. Many Hindu authorities claim that regular temple worship to one of the deities of the devotional cults procures the same results for the worshiper as did the performance of one of the great Vedic sacrifices, and one who provides the patronage for the construction of a temple is called a "sacrificer" (*yajamāna*).

Temples. The erection of a temple, which belongs to whoever paid for it or to the community that occupies it, is a meritorious deed recommended to anyone desirous of heavenly reward. The choice of a site, which should be serene and lovely, is determined by astrology and divination as well as by its location with respect to human dwellings; for example, a sanctuary of a benevolent deity should face the village. The construction of a temple is, because of its symbolic value, described in great detail. There is much diversity in size and artistic value, ranging from small village shrines with simple statuettes to great temple-cities whose boundary walls, pierced by monumental gates (gopura), enclose various buildings, courtyards, pools for ceremonial bathing, and sometimes even schools, hospitals, and monasteries. From the point of view of construction there is no striking difference between Śaiva and Vaishnava sanctuaries, which are easily recognizable by the image or symbols in the centre, the images on the walls, the symbol fixed on the finial (crowning ornament) of the top, and Śiva's bull, Nandi, or Vishnu's bird, Garuḍa (the theriomorphic duplicate manifestations of each god's nature), in front of the entrance. Services, which may be held by any qualified member of the community, are neither collective nor carried out at fixed times. Those present experience, as spectators, the fortifying and beneficial influence radiating from the sacred acts. Sometimes worshipers assemble to meditate, to take part in chanting, or to listen to an exposition of doctrine. The puja (worship) performed in public "for the well-being of the world" is, though sometimes more elaborate, largely identical with that executed for personal interest. There are, on the other hand, many regional differences, and even significant variations within the same community.

Puja. Hindu worship (puja) consists essentially of an invocation, a reception, and the entertainment of God as a royal guest. It normally consists of 16 "attendances" (*upacāra*): invocation by which the omnipresent God is invited to direct his attention to the particular worship; the offering of a seat, water (for washing the feet, for washing the hands, and for rinsing the mouth), a bath, a garment, a sacred thread, perfumes, flowers, incense, a lamp, food, and homage; and a circumambulation of the image and dismissal by God.

The Pāñcarātra Vaishnavas in South India introduced the songs of the Dravidian poets into their temple cult and regard these poets and their great teachers as incarnations of God, even to the point of worshiping their images. The Śaivas also have songs of their own but were, generally speaking, more open to Tantric elements and to the admission in their cult of dances executed by dancing girls. In both religious groups, some communities cling to the traditional Sanskrit mantras while others also use other languages.

The significance of the temple door

The first phase of worship is the reverential opening of the temple door and the adoration of the powers presiding over it: according to the Vaikhānasa Vaishnavas, the symbolic opening of heaven; and to the Śaivas, an act to secure the building's protection. The divine powers whose images are carved in the doorjambs promote the process of transmutation without which man cannot even enter into the presence of God, whose image is established in the cella (*garbhagṛha*). This image is honoured with gifts, notably flowers, fruit, and perfumes. Small portions of the consecrated food (*prasāda*) are given to visiting worshipers. The offering into the fire (*homa*) of Vedic origin has been retained in nearly all extended puja ceremonies. The main purpose of the rites is the meditative identification of the worshiper with the divine Presence; the enactment, in a gradual process of development, of the realization of the union of the worshiper's soul and God. The Vaikhānasas distinguish between the transcen-

Photo Atlas-Oumic, Paris



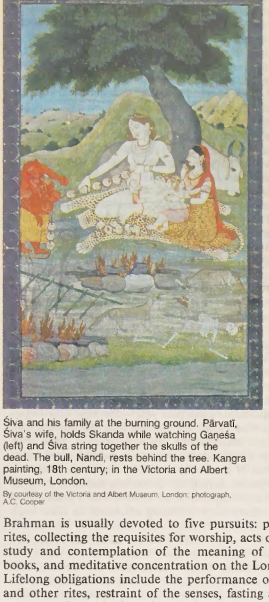
A temple puja to Krishna.

dent and unanalyzable Brahman and its immanent and analyzable aspect and invoke God to descend out of compassion from the immovable image—the permanent "seat" of the former—into a movable cult image in which he converses with the world, represented by the worshiper. Those denominations (both Śrīvaiṣṇavas and Śaivas) that adopted Tantric practices believe that God comes, during these ceremonies, also out of the worshiper's heart or that the worshiper's soul leaves his body to reach God's feet in heaven, to descend from there in a new body that is meditatively created.

A remarkable rite of yogic-Tantric origin, also used in other ritual contexts, is the transmutation of water into the elixir of life and immortality (*amṛta*), the essential element of which is drawn from the spot between the worshiper's eyebrows, regarded as the seat of Śiva's highest aspect.

Śaivas transform themselves into Śiva by means of complicated preparatory rites, because, they say, "Śiva alone can worship Śiva." Some authorities also enjoin a mental worship and sacrifice, without which "exterior" rites are rendered senseless. The merit of the performances is often said to be entrusted to God's keeping for the sake of the worshiper. Many Vaishnavas emphasize that puja is meant to propitiate God disinterestedly.

Śaiva rites. Ascetic tendencies were much in evidence among the Pāśupatas, the oldest Śaiva tradition in North India, the last adherents of which now live in Nepal. Pāśupatas often gave offense because of their customs and ritual practices. Their yoga, consisting of a constant meditative contact with God in solitude, required that they frequent burning places for cremated bodies. More extreme groups carried human skulls (hence the name Kāpālikas, from *kapāla*, "skull") which they used as bowls for liquor into which they projected and worshiped Śiva as Kāpālika, "the Skull Bearer," or Bhairava, "the Frightful One," and then drank to become intoxicated. Their belief was that an ostentatious indifference to anything worldly was the best method of severing the ties of samsara.



Śiva and his family at the burning ground. Pārvatī, Śiva's wife, holds Skanda while watching Gaṇeśa (left) and Śiva string together the skulls of the dead. The bull, Nandi, rests behind the tree. Kangra painting, 18th century; in the Victoria and Albert Museum, London.

By courtesy of the Victoria and Albert Museum, London: photograph, A.C. Cooper

Vīraśaivas (Liṅgāyats)

The view and way of life peculiar to the Vīraśaivas, or Liṅgāyats (liṅgam-bearers), in southwestern India is mainly characterized by a deviation from some common Hindu traditions and institutions such as sacrificial rites, temple worship, pilgrimages, child marriages, and inequality of the sexes. Initiation (*dīkṣā*) is, on the other hand, an obligation laid on every member of the community. The spiritual power of the guru is bestowed upon the newborn and converts, who receive the eightfold shield, which protects devotees from ignorance of the supremacy of God and guides them to final beatitude, and the liṅgam (phallic symbol). The miniature liṅgam, the centre and basis of all their religious practices and observances, which they always bear on their body, is God himself concretely represented. Worship is due it twice or three times a day. When a Liṅgāyat "is absorbed into the liṅgam" (*i.e.*, dies), his body is not cremated, as is customary in Hinduism, but is interred, like ascetics of other groups. Those Liṅgāyats who have reached a certain level of holiness are supposed to die in the state of emancipation.

Varieties of Śaivism

Śaivism, though inclined in doctrinal matters to adoptive inclusivism, inculcates some fundamental lines of conduct: one should worship one's spiritual preceptor (guru) as God himself, follow his path, consider him to be present in oneself, and dissociate oneself from all opinions and practices that are incompatible with the Śaiva creed. Yet some of Śiva's devotees also worship other gods, and the "Śivaization" of various ancient traditions is sometimes rather superficial. Like many other Indian religions, the Śaiva-siddhānta has developed an elaborate system of ethical philosophy, primarily with a view to preparing the way for those who aspire to liberation. Because dharma leads to happiness, there is no distinction between sacred and secular duties. All deeds are performed as services to God and with the conviction that all life is sacred and God-centred. A devout way of living and a nonemotional mysticism are thus much recommended. Kashmir Śaivism developed the practice of a simple method of salvation: by the recognition (*pratyabhijñā*)—direct, spontaneous, technique-free, but full of bhakti—of one's identity with God.

Vaishnava rites. The day of the faithful Śrīvaiṣṇava Brahman is usually devoted to five pursuits: purificatory rites, collecting the requisites for worship, acts of worship, study and contemplation of the meaning of the sacred books, and meditative concentration on the Lord's image. Lifelong obligations include the performance of sacrifices and other rites, restraint of the senses, fasting and soberness, worship, recitation of the scriptures, and visits to sacred places. In addition, to those who aspire to liberation, Rāmānuja recommends concentration on God, a virtuous way of living, and insensibility to luck and misfortune. According to Madhva (*c.* 1199–*c.* 1278), a faithful observance of all regulations of daily conduct—including bathing, breath control, etc.—will contribute to eventual success in the quest for liberation. Devout Vaishnavas are inclined to emphasize God's omnipotence and the far-reaching effects of his grace. They attach much value to the repeated murmuring of his name or sacred formulas (*japa*) and to the praise and commemoration of his deeds as a means of self-realization and of unification with his essence. Special stress is laid on ahimsa as a virtue.

SACRED TIMES AND PLACES

Festivals. Hindu festivals are combinations of religious ceremonies, semi-ritual spectacles, worship, prayer, lustrations, processions (to set something sacred in motion and to extend its power throughout a certain region), music, dances (which by their rhythm have a compelling force), magical acts—participants throw fertilizing water or, during the Holī festival, coloured powder at each other—eating, drinking, licentiousness, feeding the poor, and other activities of a religious or traditional character. The original functions of these activities are clear from ancient literature and anthropological research: they are intended to purify, avert malicious influences, renew society, bridge over critical moments, and stimulate or resuscitate the vital powers of nature (hence the term *utsava*, meaning both the generation of power and a festival). Because such festivals relate to the cyclical life of nature, they are supposed to prevent it from stagnating. These cyclic festivals—which may last for many days—continue to be celebrated throughout India.

Such festivals refresh the mood of the participants, further the consciousness of their own power, and help to compensate for their sensations of fear and inferiority concerning the unknown forces of nature. Such mixtures of worship and pleasure require the participation of the entire community and create harmony among its members, even if not all participants are now aware of the original character of the festival. There are also innumerable festivities in honour of specific gods, celebrated by individual temples, villages, and religious communities.

An important festival, formerly celebrating Kāma, the god of sexual desire, survives in the Holī, a saturnalia connected with the spring equinox and in western India with the wheat harvest. The lower classes observe it in its boisterous and licentious form. There are local variants: among the Marāṭhās, heroes who died on the battlefield are "danced" by their descendants, sword in hand, until they believe themselves possessed by the spirits of the heroes. In Bengal, swings are made for Krishna; in other regions a bonfire is also essential. The mythical tradition that accounts for the festival describes how young Prahlāda, in spite of his demonic father's opposition, persisted in worshiping Vishnu and was carried into the fire by the female demon Holikā, the embodiment of evil, who herself was believed to be immune to the ravages of fire. Through Vishnu's intervention, Prahlāda emerged unharmed, while Holikā was burned to ashes. The bonfires are intended to commemorate this event or rather to reiterate the triumph of virtue and religion over evil and sacrilege. This explains why objects representing the sickness and impurities of the past year—the new year begins immediately after Holī—are thrown into the bonfire, and it is considered inauspicious not to look at it. Moreover, people pay or forgive debts, reconcile quarrels, and try to rid themselves of the evils, conflicts, and impurities that have accumulated during the preceding months, translating the central conception of the festival into a justification for dealing anew with continuing situations in their lives. The New Year festival, according to another Indian calendar, Dīwālī, though celebrated by all classes of society, is traditionally believed to have been given by Vishnu to the Vaiśyas (traders, etc.); it takes place in October, with worship and ceremonial lights in honour of Lakṣmī, the goddess of wealth and good fortune; fireworks to chase away the spirits of the deceased; and gambling, an old ritual custom intended to secure luck for the coming year. The nine-day Durgā festival, or Navarātrī, is, especially in Bengal, splendid homage to Śakti, and in South India, a celebration of Rāma's victory over Rāvaṇa.

New Year festivals

Pilgrimages and fairs. Like processions, pilgrimages (*tīrthayātrā*) to holy rivers (*tīrtha*) and other places were already known in Vedic and epic times and are even now one of the most remarkable aspects of Indian religious life. Many sections of the *Purāṇas* eulogize temples and the sacredness of places situated in beautiful scenery or wild solitude (especially the Himalayas). The whole of India, and especially Kurukshetra (presumed to be the scene of the great war portrayed in the *Mahābhārata*) in the northwest, is considered holy ground that offers everyone the opportunity to reach emancipation. The number of places of pilgrimage of regional significance amounts to many hundreds, but some of them (Ayodhyā, Mathura, Hardwār, Vārānasi [Benares], Kānchipuran, Ujjain, and Dwārka) have for many centuries possessed exceptional holiness. The reason for such sanctity derives from their location on the bank of a holy river, especially of the Ganges, from their connection with legendary figures of antiquity who are said to have lived there, or from the local legend of a manifestation of a god. Many places are sacred to a specific god; the district of Mathura, for example, encompasses many places of pilgrimage connected with the Krishna legends. Visits to holy places may bestow special benefits upon pilgrims; temples or ponds dedicated to Sūrya (the Sun) are visited in order to recover from leprosy, other places to escape from astrological threats. Pilgrimages to Gayā (Bihār state)—where visitors are escorted around the sacred centres by Brahman temple priests who maintain certain ritual connections with their clients—are undertaken for the sake of the welfare of deceased ancestors. In most cases, however, the devotee hopes for worldly rewards (health, wealth, children) or for spiritual rewards such as deliverance from sin or pollution, preservation of religious merit, rebirth in a heaven, or even emancipation. The last prospect is held out to those who, when death is near, travel to Vārānasi to die near the Ganges.

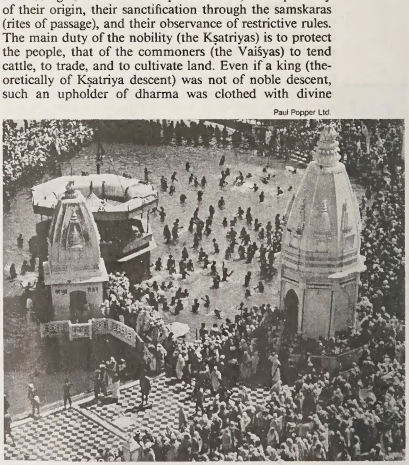
On special occasions, be they auspicious or, like a solar eclipse, inauspicious, the devout crowds increase enormously. Most important shrines also organize gatherings (melas), that are partly fairs, partly religious demonstrations. These journeys, which are undertaken by individuals or groups in order to discharge a vow or to please a god, confirm the devotees in their faith, provide them with an opportunity for spiritual retreat, or bring their inner life nearer to a state of perfection. They have contributed much to the spread of religious ideas and the cultural unification of India.

RITUAL AND SOCIAL STATUS

Some observers have claimed that Hinduism is as much a way of social life as it is a religion. The caste system, which has organized Indian society for many millennia, is thoroughly legitimated by and intertwined with Hindu religious doctrine and practice. Four social classes, or varnas—Brahmans, Kṣatriyas, Vaiśyas, and Śūdras—provide the simplified structure for the enormously complicated system of thousands of castes and subcastes within Indian society. Although it is not certain whether a society limited to four classes was ever more than a theoretical ideal, there is a sense in which they map out socioreligious reality. Such is evident from the Puruṣa hymn (Rigveda 10.90), in which the statement that the Brahman was the Puruṣa's mouth, the nobleman (Kṣatriya) his arms, the Vaiśya his thighs, and the Śūdra his feet, gives an idea of their functions and mutual relations.

The Brahmans, whatever their worldly avocations, claim to be by virtue of their birth a perpetual incarnation of the dharma, guardians and dispensers of divine power, entitled to teach the Veda, sacrificing for others and accepting gifts and subsistence; the term alms is misleading, and the *dakṣiṇā* offered at the end of a rite to a Brahman officiant is not a fee but an oblation through which the rite is made complete. Brahmans are held to be the highest of all human beings because of their preeminence, the superiority of their origin, their sanctification through the samskaras (rites of passage), and their observance of restrictive rules. The main duty of the nobility (the Kṣatriyas) is to protect the people, that of the commoners (the Vaiśyas) to tend cattle, to trade, and to cultivate land. Even if a king (theoretically of Kṣatriya descent) was not of noble descent, such an upholder of dharma was clothed with divine

Paul Popper Ltd.



Pilgrims bathing in the Ganges River at Hardwār, India.

authority. He was consecrated by means of a complex and highly significant ritual; he was Indra and other gods (deva) incarnate. The emblems or paraphernalia of his office represent sovereign authority; the white umbrella of state is, for example, the residence of Śrī-Lakṣmī, the goddess of fortune. All three higher classes, claiming Aryan descent, had to sacrifice and to study the Veda, although the responsibilities of the Vaiśyas in sacred matters were less demanding.

While this tripartition seems, in the main, to have been inherited from Indo-European times, the fourth class (the Śūdras), whose sole duty it was "to serve meekly" (*Mānava Dharmaśāstra* 1.91) the other classes, are partly descended from the subjugated non-Aryans, a fact that accounts for their many disabilities and exclusion from religious status. According to Hindu tradition, the Veda should not be studied in their presence, but they may listen to the recitation of epics and *Purāṇas*. They are permitted to perform the five main acts of worship (without Vedic mantras) and undertake observances, but even today they maintain various ceremonies of their own, carried out without Brahmanic assistance. Yet a distinction is often made among Śūdras. Some are purer and have a more correct behaviour and way of living than others, the former tending to assimilate with higher castes, the latter to rank with the lowest in the social scale, who, often called *cāṇḍālas*, were at an early date sweepers, bearers of corpses, or charged with other impure occupations. Ritual purity was indeed an important criterion; impure conduct and neglect of Veda study and the rules regarding forbidden food might suffice to stigmatize a "twice-born man" as a Śūdra. On the other hand, in later times the trend of many communities has been toward integrating all Śūdras into the Brahmanic system. The Brahmans, who have far into modern times remained, on the whole, a respected, traditional, and sometimes intellectual upper class, were generally (until the 1930s) much in demand because of their knowledge of rites and traditions. Although Kṣatriya rank is claimed by many whose title is one of function or creation rather than of inheritance, this class is now rare in many regions. Moreover, for a considerable time none of the four varnas represented anything other than a series of hierarchically arranged groups of castes.

Castes. The origin of the caste system is not known with certainty. Hindus account for the proliferation of the castes (*jātis*, literally "births") by the subdividing of the four classes, or varnas, due to intermarriage (which is prohibited in Hindu works on dharma). Modern theorists, however, tend to assume that castes arose from differences in family ritual practices, racial distinctions, and occupational differentiation and specialization. Many modern scholars also doubt whether the simple varna system was ever more than a theoretical socioreligious ideal and have emphasized that the highly complex division of Hindu society into nearly 3,000 castes and subcastes was probably in place even in ancient times.

In general, a caste is an endogamous hereditary group of families, bearing a common name; often claiming a common descent; as a rule professing to follow the same hereditary calling; clinging to the same customs, especially regarding purity, meals, and marriages; and often further divided into smaller endogamous circles. Moreover, tribes, guilds, or religious communities characterized by particular customs—for example, the Liṅgāyats—could easily be regarded as castes. The status of castes varies in different localities. Although social mobility is possible, the mutual relationship of castes is hierarchically determined: local Brahman groups occupy the highest place, and differences in ritual purity are the main criteria of position in the hierarchy. Most impure are the untouchables, or, to use modern names, the exterior or scheduled castes, which, however, have among themselves numerous divisions, each of which regards itself as superior to others.

Traditional Hindus are inclined to emphasize that the ritual impurity and "untouchability" inherent in these groups does not essentially differ from that temporarily proper to mourners or menstruating women. This, and the fact that some exterior group or other might rise in estimation and become an interior one, or that individual outcastes might be well-to-do, does not alter the fact that the spirit of exclusiveness was in the course of time carried to extremes. The scheduled castes were subjected to various socioreligious disabilities before mitigating tendencies helped bring about reform; after independence, social discrimination was prohibited, and the practice of untouchability was made a punishable offense (it was not abolished, however). Scheduled castes were barred from the use of temples and other religious institutions and from public schools. These groups also had many disabilities in relations with private persons. From the traditional Hindu point of view, this social system is the necessary complement of the principles of dharma, karma, and samsara. Corresponding to hells and heavenly regions in the hereafter, the castes are the mundane, social frame within which karma is manifested. A low social status is the inevitable result of sins in a former life but can, by virtue and merit, be followed by a better position in the next existence.

Religious orders and holy men. Those members of the various denominations who abandon all worldly attachment enter an "inner circle" or "order" that, seeking a life of devotion, adopts or develops particular vows and observances, a common cult, and some form of initiation.

Initiation. Generally speaking, Hindus are free to join an order or inner circle, and once they have joined it they must submit to its rites and way of living. The initiation (*dīkṣā*), a sort of purification or consecration involving a transformation of the aspirant's personality, is regarded as a complement to, or even a substitute for, the previous initiation ceremony (the *upanayana* that all twice-born Hindus undergo at adolescence), which it strikingly resembles. Such religious groups integrate ancient, widespread ideas and customs of initiation into the framework of either the Vaishnava or Śaiva patterns of Hinduism. Vaishnavism emphasizes their character as an introduction to a life of devotion and as an entrance into closer contact with God, although happiness, knowledge, a long life, and a prospect of freedom from karma are also among the ideals to which they aspire. Śaivas are convinced of the absolute necessity of initiation for anyone desiring final liberation and require an initiation in accordance with their rituals. All communities agree that the authority to initiate belongs only to a qualified spiritual guide (guru), usually a Brahman, who has previously received the special *guru-dīkṣā* (initiation as a teacher) and is often regarded as representing God himself. The postulant is sometimes committed to a probationary period, to training in yoga mysticism, or to instruction in the esoteric meaning of the scriptures. The initiate receives a devotional name and is given the distinctive mantras of the community, which, because they are sacred, must never be misused.

Dīkṣā

There are many complicated forms of initiation: the Vaishnavas differentiate between the members of the four classes; the Śaivas and Tantrists take into account the natural aptitude and competency of the recipients and distinguish between first-grade initiates, who obtain access to God, and higher-grade initiates, who remain in a state of holiness.

Yoga. The initiate guided by his guru may apply himself to yoga (a "methodic exertion" of body and mind) in order to attain, through mortification, concentration, and meditation, a higher state of consciousness in which he may find the supreme knowledge, achieve spiritual autonomy, and realize his oneness with the Highest (or however the ultimate goal is conceived). Yoga may be atheistic or combined with various philosophical or religious currents. Every denomination attempted to implement yogic practices on a theoretical basis derived from its own teachings. There are many different forms of yoga, and the practices vary according to the stage of advancement of the adepts. All serious yogis, however, agree in disapproving the use of yogic methods for worldly purposes.

Sectarian symbols. The typical Hindu ascetic (sadhu) usually wears a distinctive mark (*puṇḍra*) on his forehead and often carries some symbol of his religion.

If he is a Vaishnava he might possess a discus (chakra) and a conch shell (*saṅkha*), replicas of Vishnu's flaming weapon and his instrument of beneficent power and

Symbols of Vishnu and Śiva

omnipresent protection, or a *śālagrāma* stone or a tulasi plant, which represent, respectively, Vishnu's essence and that of his spouse Lakṣmi. If he is a Śaiva, he might impersonate Śiva and carry a trident (*triśūla*), denoting empire and the irresistible force of transcendental reality; wear a small lingam; carry a human skull, showing that he is beyond the terror inspired by the transitoriness of the world; or smear his body with apotropaic (supposed to avert evil) and consecratory ashes. These emblems are sacred objects of worship because the divine presence, when invoked by mantras, is felt to be in them.

The attitude toward asceticism has always been ambivalent. On the one hand, there is a genuine regard for hermits and wandering ascetics and a desire to gain spiritual merit by feeding religious mendicants. On the other hand, the fact that fringe members of society may find a sort of respectable status among Śaiva ascetics often led to a decline in the moral reputation of the latter. (B.K.S./Ed.)

Cultural expressions: visual arts, theatre, and dance

The structure of Indian temples, the outward form of images, and indeed the very character of Indian art are largely determined by religion and a traditional view of the world, which penetrated the other provinces of culture and welded them into a homogeneous whole.

Indian art is highly symbolic. The much-developed ritual-religious symbolism presupposes the existence of a spiritual reality that, being in constant touch with phenomenal reality, may make its presence and influence felt and can also be approached through the symbols that belong to both spheres.

Mental and external fashioning of images

The production of objects of symbolic value is therefore more than a technique. The artisan must model a cult image after the ideal prototype that appears in his mind (in certain canonical forms) only when he has brought himself to a state of supranormal consciousness. After undergoing a process of spiritual transformation himself, he also transforms the material of which the image is to be made into a receptacle of divine power. Like the artisan, the worshiper (*sādhaka,* "the one who wishes to attain the goal") must grasp the esoteric meaning of a statue, picture, or pot and identify his or her self with the power residing in it. The usual offering, a handful of flowers, is the vehicle used to convey the worshiper's "life-breath" into the external image, which has already been transformed into an adequate internal vision of the same divine power.

TYPES OF SYMBOLS

If they know how to handle the symbols, the worshipers—who must achieve their object themselves and cannot come into contact with God unless they insistently invoke him—have at their disposition an instrument for utilizing the possibilities lying in the depths of their own subconscious as well as a key to the mysteries of the forces dominating the world.

Yantra and mandala. The general term for an "instrument [for controlling]" is yantra, which, while denoting in a wider sense cult images, pictures, and other such aids to worship, is often especially applied to ritual diagrams. Any yantra represents some aspect of the divine and enables devotees to worship it immediately within their hearts while identifying themselves with it. Except in its greater linear complication, a mandala does not differ from a yantra and both are drawn during a highly complex ritual in a purified and ritually consecrated place. The meaning and the use of both are similar, and they may be permanent or provisional. A mandala, delineating a consecrated place and protecting it against disintegrating forces represented in demoniac cycles, is the geometric projection of the universe, spatially and temporally reduced to its essential plan. It represents in a schematic form the whole drama of disintegration and reintegration, and the adept can use it to identify with the forces governing these. As in temple ritual, a vase is employed to receive the divine power so that it can be projected into the drawing and then into the person of the adept. Thus the mandala becomes a support for meditation, an instrument to provoke visions of the unseen. A good example of a mandala is the *śrīcakra,* "the Wheel of Śrī" (*i.e.,* of God's shakti) composed of four isosceles triangles with the apices upward, symbolizing Śiva, and five isosceles triangles with the apices downward, symbolizing Śakti; the nine triangles are of various sizes and intersect with one another. In the middle is the power point (*bindu*), visualizing the highest, the invisible, elusive centre from which the entire figure and the cosmos expand. The triangles are enclosed by two rows of (eight and 16) petals, representing the lotus of creation and reproductive vital force. The broken lines of the outer frame denote the figure to be a sanctuary with four openings to the regions of the universe. A "spiritual" foundation is provided by a yantra, called the mandala of the Puruṣa (spirit) of the site, that is also drawn on the site on which a temple is built. This rite is a reenactment of a variant of the myth of Puruṣa, an immortal primeval being who obstructed both worlds until he was subdued by the gods; the parts of his body became the spirits of the site.

Śrīcakra mandala

Lingam and yoni. One of the most common objects of worship, whether in temples or in the household cult, is the lingam (phallus). Often much stylized and an austere rather than literally sexual symbol, erect and representing the cosmic pillar, it emanates its all-producing energy to the four quarters of the universe. As the symbol of male creative energy it is frequently combined with its female counterpart (yoni), the latter forming the base from which the lingam rises. Although the lingam originally may have had no relation to Śiva, it has from ancient times been regarded as symbolizing Śiva's creative energy and is widely worshiped as his fundamental form.

Visual theology in icons. The beauty of cult objects contributes to their force as sacred instruments: their ornamentation facilitates the process of inviting the divine power into them. Statues of gods are not intended to imitate ideal human forms but to express the supernatural. A divine figure is a "likeness" (*pratimā*), a temporary benevolent or terrifying expression of some aspect of a god's nature. Iconographic handbooks attach great importance to the ideology behind images and reveal, for example, that Vishnu's eight arms stand for the four cardinal and intermediate points of the compass and that his four faces, illustrating the concept of God's fourfoldness, typify his strength, knowledge, lordship, and potency. The emblems express the qualities of their bearers—*e.g.,* a deadly weapon symbolizes destructive force, many-headedness omniscience. Much use is made of gestures (mudras), conventional devices for denoting activities that express an idea; thus, the raised right hand, in the "fear-not" gesture (*abhaya-mudrā*), bestows protection. Every iconographic detail has its own symbolic value, helping devotees to direct their energy to a deeper understanding of the various aspects of the divine and to proceed from external to internal worship. For many Indians, an installed and consecrated image becomes a container of concentrated divine energy; according to Hindu theists, it is an instrument for ennobling the worshiper who realizes God's presence in it.

Purposes of Hindu iconography

THE ARTS

Religious principles in sculpture and painting. The dance executed by Śiva as king of dancers (Naṭarāja), the visible symbol of the rhythm of the universe, represents God's five activities: he unfolds the universe out of the drum held in one of his right hands; he preserves it by uplifting his other right hand in *abhaya-mudrā;* he reabsorbs it with his upper left hand, which bears a tongue of flame; his transcendental essence is hidden behind the garb of apparitions, and grace is bestowed and release made visible by the foot that is held aloft and to which the hands are made to point; and the other foot, planted on the ground, gives an abode to the tired souls struggling in samsara. Another dance pose adopted by Śiva is the doomsday tandava, executed in his destructive Bhairava manifestation, usually with 10 arms and accompanied by Devī and demons. The related myth is that Śiva conquered a mighty elephant demon whom he forced to dance until he fell dead; then, wrapped in the blood-dripping skin of his victim, the god executed a horrendous dance of victory.

Images sustain the presence of the god: when Devī is

shown advancing against the buffalo demon, seated on her lion, she represents the affirmative forces of the universe and the triumph of divine power over wickedness. Male and female figures in uninterrupted embrace, as in Śaiva iconography, signify the union of opposites and the eternal process of generation. Lovers sculpted on temples are auspicious symbols on a par with foliage, water jars, and other representatives of fertility.

Like literature and the performing arts, the visual arts also contributed to the perpetuation of myths. Hindu sculpture tends to be less narrative than Buddhist, which delights in scenes from the Buddha's lives. In Hindu sculpture the tendency is toward hieratic poses of a god in a particular conventional stance (*mūrti*), which, once fixed, perpetuates itself. An icon is a frozen incident of a myth. For example, one *mūrti* (image) of Śiva is the "destruction of the elephant," in which Śiva appears dancing before and below a bloody elephant skin that he holds up before the image of his horrified consort; the stance is the summary of his triumph over the elephant demon. A god may also appear in a characteristic pose while holding in his multitudinous hands his various emblems, on each of which hangs a story. Carvings, such as those that appear on temple chariots, tend to be more narrative; even more so are the miniature paintings of the Middle Ages. A favourite theme in the latter is the myth of the cowherd god Krishna and his love of the cowherd wives (*gopīs*).

Religious organization of sacred architecture. Temples must be erected on a site that is *śubha* (*i.e.,* suitable, beautiful, auspicious, and near water) because the gods will not come to other places. However, temples are not necessarily designed to be congenial to their surroundings, because a manifestation of the sacred is an irruption, a break in phenomenal continuity. Temples are said to constitute an opening in the upward direction to ensure communication with the gods; they are visible representations of a cosmic pillar and their site is said to be a navel of the world. Their outward appearance must raise the expectation of meeting with God. Their erection is a reconstruction and reintegration of Puruṣa-Prajāpati, enabling him to continue his creative activity, and the finished monuments are symbols of the universe that is the unfolded One. The owner of the temple (*i.e.,* the individual or community that paid for its construction)—also called the sacrificer—participates in the process of reintegration and experiences his spiritual rebirth in the small cella, aptly called the "womb room" (*garbhagṛha*), by means of meditative contact with God's presence, symbolized or actualized in his consecrated image. The cella is in the centre of the temple above the navel—*i.e.,* the foundation stone; it may contain a jar filled with the creative power (shakti) that is identified with the goddess Earth (who bears and protects the monument), three lotus flowers, and three tortoises (of stone, silver, and gold) that represent Earth, atmosphere, and heaven. The tortoise is a manifestation of Vishnu bearing the cosmic pillar; the lotus is the symbol of the expansion of generative possibilities. The vertical axis or tube (coinciding with the cosmic pillar), which connects all parts of the building and is continued in the finial on the top, corresponds with the mystical vertical vein in the body of the worshiper through which his soul rises to unite itself with the Highest.

Requirements of temple architecture and planning

The designing of Hindu temples, like that of religious images, was codified in the *Śilpa-śāstras* (craft textbooks), and every aspect of the design was believed to be symbolic of some feature of the cosmos. The idea of microcosmic symbolism is strong in Hinduism and comes from Vedic times; the Brāhmaṇa texts are replete with similar cosmic interpretations of the many features of the sacrifice. This same Vedic idea of the correspondence (*bandhu*) between microcosm and macrocosm was applied to the medieval temple, which was laid out geometrically to mirror the structure of the universe, with its four geometric quarters and a celestial roof. The temple also represents the mountain at the navel of the world and often somewhat resembles a mountain. On the periphery were carved the most worldly and diverse scenes, including luxurious celebrations of human life: battle scenes, hunts, circuses, animals, birds, as well as images of the gods.

Temple design as cosmic symbolism

The erotic scenes carved at Khajurāho in Madhya Pradesh and Konārak in Orissa express a general exuberance that may be an offering of thanksgiving to the gods who created all. However, that same swarming luxuriance of life in all of its aspects may also reflect the concern that one must set aside worldly temptations upon the threshhold of the sacred space of the temple, for the carvings only decorate the outside of the temple; at the centre, the sanctum sanctorum, there is little if any ornamentation, except for a stark symbol of the god or goddess. Thus, these carvings simultaneously express a celebration of samsara and a movement toward moksha.

Theatre and dance. Theatrical performances are also events that can be used to secure blessings and happiness; the element of recreation is indissolubly blended with edification and spiritual elevation. The structure and character of the classical Indian drama reveal its origin and function: it developed from the last part of a magico-religious ceremony, which survives as a ritual introduction, and begins and closes with benedictions. Drama is produced for festive occasions with a view to spiritual and religious success (*siddhi*), which must also be prompted by appropriate behaviour from the spectators; there must be a happy ending; the themes are borrowed from epic and legendary history; the development and unraveling of the plot are retarded; and the envy of malign influences is averted by the almost obligatory buffoon (*vidūṣaka,* "the spoiler"). There are also, in addition to films, which often use the same religious and mythic themes, *yātrās*, a combination of stage play and various festivities that have contributed much to the spread of the Purāṇic view of life.

Yātrās

By courtesy of the Gujarat Museum Society, Ahmadabad, India



Krishna dancing with the *gopīs*, painting from western Rājasthān, c. 1610. In the N.C. Mehta Collection of the Gujarāt Museum Society, Ahmadābād, India.

Dancing is not only an aesthetic pursuit but also a divine service. Hence there are halls for sacred dances annexed to some temples. The rhythmic movement has a compelling force, generating and concentrating power or releasing superfluous energy. It induces the experience of the divine and transforms the dancer into whatever he or she impersonates. Thus, many tribal dances consist of symbolic enactments of events (harvest, battles) in the hope that they will be accomplished successfully. Musicians and dancing girls accompany processions to expel the demons of cholera or cattle plague. Even today, religious themes and the various relations between humans and God are danced and made visual by the codified symbolic meanings of gestures and movements (see SOUTH ASIAN ARTS).

(A.L.B./J.A.B.v.B./W.Do./Ed.)

The place of Hinduism in world religions

HINDUISM AND OTHER RELIGIONS OF INDIAN ORIGIN

Hinduism, Buddhism, and Jainism originated out of the same milieu: the circles of world-renouncers of the 6th century BC. Although all share certain non-Vedic practices (such as renunciation itself and various yogic meditational techniques) and doctrines (such as the belief in rebirth and the goal of liberation from perpetual transmigration), they differ in the respect they accord to the Vedic tradition.



Vishnu with his 10 avatars (incarnations): Fish, Tortoise, Boar, Man-Lion, Dwarf, Rāma-with-the-Ax, King Rāma, Krishna, Buddha, and Kalkin. Painting from Jaipur, India, 19th century; in the Victoria and Albert Museum, London.
By courtesy of the Victoria and Albert Museum, London

Virtually all Hindus affirm the sacredness and authority of the Veda; Buddhists and Jains do not and therefore are regarded as less than orthodox by Hindus.

Buddhism. Although Buddhism did not interfere with Hindu customs and usages, allowing its adherents to approach Hindu or local supernatural powers for immediate favours, Hindu criticism of Buddhism came mainly from Brahman philosophers who opposed its adherents because they rejected the authority of the Veda and the Brahmans and the doctrine of the atman (soul) and because they admitted persons of any age and caste to monastic life. The spread of Buddhism was often regarded as an indication of degeneration. In the course of time, the Buddha was recognized as an incarnation of Vishnu, but this was often qualified by the addition that Vishnu assumed this form to mislead and destroy the enemies of the Veda, and this avatar is rarely worshiped. Buddhist emblems also were often ascribed to Vishnu or Śiva. Some Buddhist shrines have remained partly under the supervision of Hindu ascetics and are visited by pilgrims notwithstanding their much neglected condition.

After the rise of Buddhological studies in the West and the archaeological discoveries and restorations beginning at the end of the 19th century had made Indians more aware of the Indian origin of Buddhism, the Republic of India adopted the Buddhist emperor Aśoka's lion capital, marking the place of Buddha's first teaching, as its national emblem. The Buddha jubilee in 1956 was an occasion for enthusiastic celebrations. The number of Indian Buddhists has again increased, due mainly to the conversion of persons of low social rank who hope for higher social status as Buddhists than they were afforded as Hindus.

Jainism. With Jainism, which always remained an Indian religion, Hinduism has so much in common, especially in social institutions and ritual life, that nowadays Hindus tend to consider it a Hindu sect. Many Jains also are inclined to fraternization. The points of difference—*e.g.*, a stricter ahimsa practice and the absence of sacrifices for the deceased in Jainism—do not give offense to orthodox Hindus (see BUDDHISM, THE BUDDHA AND; JAINISM).

HINDUISM AND ISLĀM

Because Islām was so different from Hinduism in creed and institutions, it was neither absorbed nor powerful enough to make India a Muslim country. The religious situation created by the presence of its numerous adherents always had explosive potentialities: Muslims do not respect bovine life and regard Hindu cult practices as objectionable idolatry. Although Indian Muslims, with few exceptions, are of native descent, they are theoretically outcastes with whom dealings must remain restricted by formal rules; however, as with Christians, they are less polluting than the Hindu lower castes. The Islāmic way of life meets with opposition, and orthodox Muslims and Hindus do not ordinarily intermarry or dine together. This situation has had acute and even devastating consequences, but it does vary somewhat from region to region, from village to village, and from class to class. Very often mutual differences are accepted. Although they repudiate caste, Muslims often observe it in practice, and some have even retained their original caste organization after their conversion to Islām.

Syncretism of Hindu and Islāmic religion

Throughout centuries of close proximity and daily interaction, Hindus and Muslims have made efforts to accommodate the existence of the other religion within their own. One manifestation of such syncretism occurred among mystically inclined groups who believed that the one God, or "the universal principle," was the same regardless of whether it was called Allah or *brahman*. Various syntheses between the two religions, including Sikhism and other movements that emphasize nonsectarianism, have arisen in North India.

Those who, like Gandhi, could not understand the intolerance of orthodox Islām sympathized with the moderation and eclecticism of such groups. Most of the educated class, however, have always remained aware of the cleavage. To the Muslims—who, as part of an ecumenical community stretching over large parts of Asia and Africa, are concerned about the political and religious crisis of Islām since the late 19th century—the collapse of the Mughals after the Indian Mutiny (1857–58) was a severe blow that worsened relations with Hindus. This is particularly true because anti-Muslim tendencies had won ground since the renascent Hinduism of the Marāṭhā movement and in later times in the Arya Samaj (see above), while Muslims became self-assertive and even more determined to maintain their distinctive position. After the partition of the subcontinent into India and Pakistan—partly based on religious differences—and independence (1947), the political controversies between India and Pakistan constituted a further complication for relations between the religions.

HINDUISM AND CHRISTIANITY

The relations between Hinduism and Christianity have been shaped by unequal balances of political power and cultural influence. Although small communities of Christians have lived in South India since the middle of the 1st millennium, Christianity was widely introduced into the Indic subcontinent only in modern times by missionaries working under the auspices of British colonialism and imperialism. Their denigration of Hindu beliefs and practices—such as image worship and widow burning—provoked a Hindu response. Beginning in the 19th century and continuing to the present, a movement that might be called neo-Vedānta has emphasized the monism of certain Upanishads, decried "popular" Hindu "degenerations" such as the worship of idols, and acted as an agent of social reform, modernization, and dialogue between other world religions.

The relations between Hindus and Christians, then, have been complicated. Many Hindus are ready to accept the ethical teachings of the Gospels, particularly the Sermon on the Mount (whose influence on Gandhi is well-known) but reject the theological superstructure. Many adherents of bhakti movements—the Christian influence on which has been grossly exaggerated—feel that the Christian conceptions, which are regarded as a kind of bhakti, do not realize in God the multiplicity of human relations of love and service. Educated Hindus, though assimilating some Christian ideas, often regard missionary propaganda as an attack on their national genius and time-honoured institutions and take offense at what they regard as the disrespectful utterances of Christian missionary literature.

They are averse to the organization, the reliance on authorities, and the exclusiveness of Islām and Christianity, considering these as obstacles to harmonious cooperation. They subscribe to Gandhi's opinion that missionaries should confine their activities to humanitarian service. Since independence, conversion has indeed been viewed with disfavour by many influential Indians, who often also find in Hinduism what might be attractive in Christianity.

The Arya Samaj movement

Movements that advocate a Hindu theism designed to rival Islām and Christianity, like the Arya Samaj, make serious efforts to reconvert Christians to the Hindu community. People tolerate the proximity of Christian converts, even if they transgress Hindu taboos, provided they form a more or less separate community. Thus Christians often form castes or endogamous bodies analogous to castes. They sometimes are even admitted to temples to which untouchable Hindus have no entrance. In Malabar, due to their high economic position, Christians came to be practically equal with Brahmans. Nationalism has challenged the more serious-minded Indian Christians to express the genius of their faith in Indian modes and patterns. This has led, since 1921, to the emergence of Christian ashrams in the south. The dialogue between Hinduism and Christianity is more or less institutionalized at Bangalore in Karnātaka state, where the Christian Institute for the Study of Religion and Society is located. Its bulletin offers an opportunity for discussion between, for example, Christians and supporters of the Ramakrishna Mission.

(B.K.S./Ed.)

BIBLIOGRAPHY. Among the many overviews of Hinduism are THOMAS J. HOPKINS, *The Hindu Religious Tradition* (1971); DAVID R. KINSLEY, *Hinduism: A Cultural Perspective* (1982); R.C. ZAEHNER, *Hinduism*, (1962, reissued 1977); and LOUIS RENOU, *Religions of Ancient India* (1953, reissued 1972). An excellent survey of all aspects of pre-Muslim ancient India is A.L. BASHAM, *The Wonder That Was India*, 3rd rev. ed. (1967, reprinted 1985). More detailed and technical accounts may be found in JAN GONDA, *Die Religionen Indiens*, 2 vol. (1960–63); and in R.C. MAJUMDAR (ed.), *The History and Culture of the Indian People*, 11 vol. (1951–69). A basic resource for the study of Hinduism is the series "The History of Indian Literature," including the volumes by JAN GONDA, *Vedic Literature (Saṃhitās and Brāhmaṇas)* (1975), and *Medieval Religious Literature in Sanskrit* (1977); and by TEUN GOUDRIAAN and SANJUKTA GUPTA, *Hindu Tantric and Śākta Literature* (1981). A detailed outline is also provided by J.N. FARQUHAR, *An Outline of the Religious Literature of India* (1920, reprinted 1967). For historical overviews, consult VINCENT A. SMITH, *The Oxford History of India*, 4th ed. (1981); ROMILA THAPAR and PERCIVAL SPEAR, *A History of India*, 2 vol. (1965–66); and D.D. KOSAMBI, *The Culture and Civilisation of Ancient India in Historical Outline* (1965, reissued 1970; U.S. title, *Ancient India: A History of Its Culture and Civilization*, 1966, reissued 1969).

For the original sources of the principal texts of Hinduism in English translation, convenient collections include AINSLIE T. EMBREE (ed.), *The Hindu Tradition* (1966, reissued 1972); WENDY DONIGER O'FLAHERTY (ed. and trans.), *Textual Sources for the Study of Hinduism* (1988); AINSLIE T. EMBREE and STEPHEN N. HAY (eds.), *Sources of Indian Tradition*, 2nd ed., 2 vol. (1988); and R.C. ZAEHNER (ed. and trans.), *Hindu Scriptures* (1966). Compendiums of Hindu mythology in translation include WENDY DONIGER O'FLAHERTY (ed. and trans), *Hindu Myths: A Sourcebook* (1975); and CORNELIA DIMMITT and J.A.B. VAN BUITENEN (eds. and trans.), *Classical Hindu Mythology: A Reader in the Sanskrit Purāṇas* (1978).

Some of the individual textual classics of Hinduism have been translated and published in the series "Sacred Books of the East": F. MAX MÜLLER and HERMANN OLDENBERG (trans.), *Vedic Hymns*, 2 vol. (1891–97, reprinted 1979), selections from the Rigveda; MAURICE BLOOMFIELD (trans.), *Hymns of the Atharva-Veda: Together with Extracts from the Ritual Books and the Commentaries* (1897, reissued 1973); JULIUS EGGELING (trans.), *The Śatapatha-Brāhmana, According to the Text of the Mādhyandina School*, 5 vol. (1882–1900, reprinted 1978); G. BÜHLER (trans.), *The Laws of Manu* (1886, reprinted 1971), with extracts from seven commentaries; JULIUS JOLLY (trans.), *The Institutes of Vishnu* (1880, reprinted 1965), and *The Minor Law-Books* (1889, reprinted 1965); GEORGE THIBAUT (trans.), *The Vedānta-Sūtras, with the Commentary by Rāmānuja*, 3 vol. (1890–1904, reprinted 1977); and HERMANN OLDENBERG (trans.), *The Grihya-Sutras: Rules of Vedic Domestic Ceremonies*, 2 vol. (1886–92, reissued 1973). See also WENDY DONIGER O'FLAHERTY (ed. and trans.), *The Rig Veda: An Anthology* (1981), a collection of 108 hymns. ROBERT ERNEST HUME (trans.), *The Thirteen Principal Upanishads*, 2nd ed. rev. (1931, reissued 1983), remains the best translation. R.C. ZAEHNER, *The Bhagavad-Gītā: With a Commentary Based on the Original Sources* (1969, reprinted 1973), is one of the best translations of this text. Translations of the epics include PRATAP CHANDRA ROY (trans.), *The Mahābharata of Krishna-Dwaipayana Vyasa*, 12 vol. (1883–96, reprinted 1981–82); J.A.B. VAN BUITENEN (ed. and trans.), *The Mahābhārata* (1973–); HARI PRASAD SHASTRI (trans.), *The Ramayana*, 3rd ed., 3 vol. (1976); and ROBERT P. GOLDMAN (trans.), *The Rāmāyaṇa of Vālmīki: An Epic of Ancient India* (1984–), with 2 vol. published by 1988. Translations of several *Purāṇas* are available in J.L. SHASTRI, *Ancient Indian Tradition & Mythology* (1970–), with 35 vol. published by 1987; and in the *Purāṇas*, ed. by ANAND SWARUP GUPTA (1968–).

Although many of the vast sources for the vernacular literatures have not been translated, there are some, including: W. DOUGLAS P. HILL (trans.), *The Holy Lake of the Acts of Rāma* (1952), a translation of the *Rāmcaritmanas* of Tulasī Dās; CH. VAUDEVILLE (trans.), *Kabir*, vol. I (1974); LINDA HESS and SHUKDEV SINGH (trans.), *The Bījak of Kabir* (1983, reissued 1986); KENNETH E. BRYANT, *Poems to the Child-God: Structures and Strategies in the Poetry of Sūrdās* (1978); JOHN STRATTON HAWLEY, *Sūr Dās: Poet, Singer, Saint* (1984); A.K. RAMANUJAN (trans.), *Speaking of Śiva* (1973), *Hymns for the Drowning: Poems for Viṣṇu* (1981), and *Poems of Love and War: From the Eight Anthologies and the Ten Long Poems of Classical Tamil* (1985); and DAVID DEAN SHULMAN, *Tamil Temple Myths: Sacrifice and Divine Marriage in the South Indian Saiva Tradition* (1980), and *The King and the Clown in South Indian Myth and Poetry* (1985).

A good introduction to tribal and Hindu folklore is provided by VERRIER ELWIN (trans.), *Tribal Myths of Orissa* (1954, reprinted 1980), *Myths of Middle India* (1949, reprinted 1977), and *Myths of the North-East Frontier of India* (1958, reissued 1968). See also BRENDA E.F. BECK *et al.* (eds.), *Folktales of India* (1987); and STUART H. BLACKBURN and A.K. RAMANUJAN (eds.), *Another Harmony: New Essays on the Folklore of India* (1986).

For the prehistoric period and the Indus Valley civilization, see WALTER A. FAIRSERVIS, JR., *The Roots of Ancient India*, 2nd ed. rev. (1975); STUART PIGGOTT, *Prehistoric India to 1000 B.C.* (1950, reissued 1962); MORTIMER WHEELER, *The Indus Civilization*, 3rd ed. (1968); and JOHN MARSHALL, *Mohenjo-Daro and the Indus Civilization*, 3 vol. (1931, reprinted 1973). The best works on the Vedic religion include the essays found in J.C. HEESTERMAN, *The Inner Conflict of Tradition: Essays in Indian Ritual* (1985); see also FRITS STAAL, C.V. SOMAYAJIPAD, and M. ITTI RAVI NAMBUDIRI, *Agni: The Vedic Ritual of the Fire Alta*, 2 vols. (1983). The classic work on the Vedic sacrifice is SYLVAIN LÉVI, *La Doctrine du sacrifice dans les Brâhmanas* (1898, reissued 1966). Still useful are ARTHUR BERRIEDALE KEITH, *The Religion and Philosophy of the Veda and Upanishads*, 2 vol. (1925, reprinted 1971); and LOUIS RENOU, *Vedic India*, trans. from French (1957, reissued 1971). Works on the relations between Vedic religion and later Hinduism are JAN GONDA, *Change and Continuity in Indian Religion* (1965); MADELEINE BIARDEAU and CHARLES MALAMOUD, *Sacrifice dans l'Inde ancienne* (1976); and BRIAN K. SMITH, *Reflections on Resemblance, Ritual, and Religion* (1989).

The literature and teachings on dharma are presented in PANDURANG VAMAN KANE, *History of Dharmaśāstra (Ancient and Mediaeval Religions and Civil Law in India)*, 2nd ed., 5 vol. in 8 (1968–77), an indispensible work. A summary of dharma is found in ROBERT LINGAT, *The Classical Law of India* (1973; originally published in French, 1967). The best study on yoga is MIRCEA ELIADE, *Yoga: Immortality and Freedom*, 2nd ed. (1969; originally published in French, 1954); also consult JEAN VARENNE, *Yoga and the Hindu Tradition* (1976; originally published in French, 1973). The doctrine of karma and rebirth as it is presented in the texts is examined in the essays collected in WENDY DONIGER O'FLAHERTY (ed.), *Karma and Rebirth in Classical Indian Traditions* (1980). CHARLES F. KEYES and E. VALENTINE DANIEL (eds.), *Karma: An Anthropological Inquiry* (1983), is also useful. Of the many works on the theoretical underpinnings of the caste system, the most influential of recent years has been the magnum opus by LOUIS DUMONT, *Homo Hierarchicus: The Caste System and Its Implications*, rev. ed. (1980; originally published in French, 1966). Also noteworthy is VEENA DAS, *Structure and Cognition: Aspects of Hindu Caste and Ritual*, 2nd ed. (1982).

For a convenient summary of the Hindu practice and ideology of image worship, consult DIANA L. ECK, *Darśan: Seeing the Divine Image in India*, 2nd rev. and enl. ed. (1985); and the essays in JOANNE PUNZO WAGHORNE, NORMAN CUTLER, and VASUDHA NARAYANAN (eds.), *Gods of Flesh/Gods of Stone: The Embodiment of Divinity in India* (1985). The best work on Hindu temples is STELLA KRAMRISCH, *The Hindu Temple*, 2 vol. (1946, reprinted 1976); see also GEORGE MICHELL, *The*

Hindu Temple: An Introduction to Its Meaning and Forms (1977). The practice of pilgrimage in general, and specifically pilgrimage to the holy city of Vārānasi, is treated in DIANA L. ECK, *Banaras: City of Light* (1982).

For an overview of the sects worshiping Vishnu or one of his forms, see JAN GONDA, *Aspects of Early Viṣṇuism* (1954, reissued 1969); SUVIRA JAISWAL, *The Origin and Development of Vaiṣṇavism: Vaiṣṇavism from 200 B.C. to A.D. 500*, 2nd rev. and enl. ed. (1981); and MILTON SINGER (ed.), *Krishna: Myths, Rites, and Attitudes* (1966, reprinted 1981). For an excellent study on Krishna, see ALF HILTEBEITEL, *The Ritual of Battle: Krishna in the Mahābhārata* (1976). Other important works include JOHN STRATTON HAWLEY, *Krishna, the Butter Thief* (1983); and DAVID R. KINSLEY, *The Divine Player: A Study of Kṛṣṇa līlā* (1979). SUSHIL KUMAR DE, *Early History of the Vaiṣṇava Faith and Movement in Bengal, from Sanskrit and Bengali Sources*, 2nd ed. (1961), is a historical approach.

For the history of Śaivism, see C.V. NARAYANA AYYAR (SADANANDA), *Origin and Early History of Śaivism in South India* (1936, reprinted 1974); and V.S. PATHAK, *History of Śaiva Cults in Northern India, from Inscriptions 700 A.D. to 1200 A.D.* (1960, reissued 1980). The mythology of Śiva is discussed in WENDY DONIGER O'FLAHERTY, *Asceticism and Eroticism in the Mythology of Śiva* (1973, reprinted as *Śiva, the Erotic Ascetic*, 1981); and in STELLA KRAMRISCH, *The Presence of Śiva* (1981).

The best overviews of Tantrism are AGEHANANDA BHARATI, *The Tantric Tradition* (1965, reprinted 1977); and EDWARD C. DIMOCK, JR., *The Place of the Hidden Moon: Erotic Mysticism in the Vaiṣṇavasahajiyā Cult of Bengal* (1966). For the worship of the goddess in her many forms, consult the essays in JOHN STRATTON HAWLEY and DONNA MARIE WULFF (eds.), *The Divine Consort: Rādhā and the Goddesses of India* (1982, reprinted 1986). See also THOMAS B. COBURN, *Devī Māhātmya: The Crystalization of the Goddess Tradition* (1984); CHEEVER MACKENZIE BROWN, *God as Mother: A Feminine Theology in India: An Historical and Theological Study of the Brahmavaivarta Purāna* (1974); DAVID KINSLEY, *Hindu Goddesses: Visions of the Divine Feminine in the Hindu Religious Tradition* (1986); and WENDY DONIGER O'FLAHERTY, *Women, Androgynes, and Other Mythical Beasts* (1980, reprinted 1982). An interesting and accessible comparison of certain themes in the worship of Krishna and the goddess is DAVID KINSLEY, *The Sword and the Flute: Kālī and Kṛṣṇa, Dark Visions of the Terrible and the Sublime in Hindu Mythology* (1975, reprinted 1977).

Emphasizing the anthropology of "popular" Hinduism are the works by LAWRENCE A. BABB, *The Divine Hierarchy: Popular Hinduism in Central India* (1975); MCKIM MARRIOTT (ed.), *Village India: Studies in the Little Community* (1955, reprinted 1986); MILTON SINGER (ed.), *Traditional India: Structure and Change* (1958, reissued 1976), and *When a Great Tradition Modernizes: An Anthropological Approach to Indian Civilization* (1972, reprinted 1980); and C.G. DIEHL, *Instrument and Purpose* (1956). A classic case history of the process known as Sanskritization is M.N. SRINIVAS, *Religion and Society Among the Coorgs of South India* (1952, reissued 1978). For a psychoanalytic approach to Hinduism, consult G. MORRIS CARSTAIRS, *The Twice-Born: A Study of a Community of High-Caste Hindus* (1957, reissued 1967); and SUDHIR KAKAR, *The Inner World: A Psycho-Analytic Study of Childhood and Society in India*, 2nd ed. rev. and enl. (1981, reprinted 1982), and *Shamans, Mystics and Doctors: A Psychological Inquiry into India and Its Healing Traditions* (1982, reissued 1984).

The standard work on the philosophical and theological aspects of various Hindu traditions is SURENDRANATH DASGUPTA, *A History of Indian Philosophy*, 5 vol. (1922–55, reprinted 1975). A fine series of volumes on Indian philosophy is KARL H. POTTER (comp.), *Encyclopedia of Indian Philosophies* (1977–), with 4 vol. published by 1987. For an analysis of one of these traditions, see JOHN BRAISTED CARMAN, *The Theology of Rāmānuja: An Essay in Interreligious Understanding* (1974).

Developments in the Hindu tradition as it confronted Western religions and modernity are covered in D.S. SARMA, *Studies in the Renaissance of Hinduism in the Nineteenth and Twentieth Centuries* (1944). Of special interest are the texts collected and translated by RICHARD FOX YOUNG, *Resistant Hinduism: Sanskrit Sources on Anti-Christian Apologetics in Early Nineteenth-Century India* (1981).

(W.Do./B.K.S.)

The Study of History

Modern historians aim to reconstruct a record of human activities and to achieve a more profound understanding of them. This conception of their task is quite recent, dating from the development in the late 18th and early 19th centuries of scientific history, cultivated largely by professional historians. It springs from an outlook that is very new in human experience: the assumption that the study of history is a natural, inevitable human activity. Before the late 18th century, historiography (the writing of history) did not stand at the centre of any civilization. History was almost never an important part of regular education, and it never claimed to provide an interpretation of human life as a whole. This was more appropriately the function of religion, of philosophy, even perhaps of poetry and other imaginative literature.

The article is divided into the following sections:

History of historiography 559
  Ancient historiography 559
    Greco-Roman era
    Early Christian era
    Early China
  Medieval historiography 562
    Europe from the 5th to the 11th century
    Europe from the 12th to the 14th century
  Byzantine historiography 564
  Muslim historiography 565
  Historiography in the European Renaissance 566
  Early modern historiography 567
  Historiography in the age of the Enlightenment 571
  Historiography in the 19th and 20th centuries 572
Methodology of historiography 573
  Source material 574
  Using source material 574
Ancillary fields 574
  Archaeology 574
    History of archaeology
    Fieldwork
    Interpretation
  Bibliography 579
    Descriptive bibliography
    Critical bibliography
  Chronology 582
    Chinese
    Japanese
    Indian
    Egyptian
    Babylonian and Assyrian
    Jewish
    Greek
    Roman
    Christian
    Muslim
    Pre-Columbian American
  Diplomatics 591
    History of the study of documents
    Diplomatic method
    Development and characteristics of chanceries
  Epigraphy 597
    Materials and techniques
    Inscriptions as historical source material
    Inscriptions as social and cultural records
    The use of inscriptions
    History of epigraphy
  Genealogy 606
    History of genealogical study
    Modern genealogy
  Paleography 609
    Types of writing materials
    Analysis of texts
  Sigillography 611
    Seals in antiquity
    Medieval European seals
    Modern use of seals
    Chinese and Japanese seals
  Textual criticism 614
    The materials of the investigation
    Critical methods
    History of textual criticism
Bibliography 620

History of historiography

ANCIENT HISTORIOGRAPHY

Greco-Roman era. The older, pre-18th-century outlook has been particularly well studied in the historiography of the ancient Greeks and Romans. But, although two of the most important ancient historians, Herodotus and Thucydides, wrote as early as the 5th century BC, when recorded Greek historiography was only just beginning, they had few successors of comparable quality. It is a symptom of the relative lack of importance attached in antiquity to this type of activity.

Ancient history was a branch of literature. The most appreciated historians were the writers who, like Thucydides, were able to touch on universal human problems or who, like the Roman author Tacitus (died *c.* AD 120), wrote in a dramatic way about important events or who, at least, attracted readers by their excellent style and skill in composition. Many of the works that lacked some of these literary qualities failed to survive.

About 1,000 ancient Greeks wrote in antiquity on historical subjects, but most of these writers are mere names. Many of the losses appear to have occurred in antiquity itself. Even historians of first rank have fared badly. Only in a few cases have complete texts of all their writings survived. Of the voluminous history of Polybius (covering originally the period 220–144 BC) only about one-third survives. Nearly half of Livy's Roman history (originally covering the period 753–9 BC) is lost. The text that remains is reasonably good only through the efforts of a group of Roman aristocrats who, in about AD 500, were trying to salvage the chief glories of Roman literature. A considerable part of Tacitus is missing, and the surviving portions of his *Annals* and *Histories* (originally AD 14–96) derive from two unique manuscripts.

Herodotus, whom the Roman statesman Cicero called "the father of history," came from the western coast of Asia Minor. The writers who preceded him were mainly Ionians from the Greek settlements in the same area. The origin of Greek historiography lies in the Ionian thought of the 6th century. The Ionian philosophers were doing something unprecedented: they were assuming that the universe is an intelligible whole and that through rational inquiries men might discover the general principles that govern it. Hecateus of Miletus, the most important Ionian predecessor of Herodotus, was applying the same critical spirit to the largely mythical Greek traditions when he wrote, early in the 5th century, "the stories of the Greeks are numerous and in my opinion ridiculous." Herodotus was more of a traditionalist, but he introduced his work as an "inquiry" (*historia*).

Egyptian and Babylonian historiography

A glance at the older historiography of the Egyptians, the Babylonians, and the other peoples of the ancient Near East will heighten one's appreciation of the novelty of the task undertaken by Herodotus. The kings of Egypt, of Babylonia and Assyria, and of the Hittites and the Persians all sought to preserve their glorious deeds for posterity in monumental inscriptions. The more im-

portant rulers also accumulated large archives, including both ordinary administrative documents and records specially commemorating their achievements. Some 20,000 clay tablets remain from the collections written for Ashurbanipal of Assyria (668–627 BC). Both in Egypt and in Babylonia lists of kings were kept in the temples, and these were sometimes supplemented by brief annals recording the principal events, though the hatred felt by certain rulers for their predecessors led to periodic destructions of older material. The exceptional meagreness of the narrative sources for Babylonian history before 747 BC seems due to the obliteration of the older annals by Nabonassar of Babylonia (ruled 747–734). Apart from changes in literary style, there was surprisingly little development over a period of more than 1,000 years in all these types of commemorative records. The inscriptions and temple records were normally intended to perpetuate the glory of the gods in whose service these rulers had accomplished great deeds. The names and dates of dynasties and of particular rulers can be reconstructed fairly adequately with the aid of these sources, but one cannot expect much accurate information about particular events. Nor, with rare exceptions, were those who had access to this material interested in using it to write continuous histories.

Herodotus and his immediate Ionian predecessors shared a very novel outlook. Its distinctive features were a lively curiosity and a capacity to treat sources in a critical spirit. Boundless curiosity about people and their diverse customs is one of the most endearing traits of Herodotus. Like other Greeks from western Asia Minor, he was particularly stimulated by contacts with the great Persian Empire, which offered opportunities for reasonably secure travel. The resultant immense widening of historical perspective is illustrated by a story told by Herodotus about Hecateus. When the latter assured the Egyptian priests at Thebes that he could trace his descent through 16 generations, the Egyptians showed him evidence of the descent of their high priests through 345 generations. Herodotus was the first to link his geographic inquiries with true history. His descriptions of the barbarian world that confronted the Greeks provided an introduction to the epic of the successful Greek resistance to the Persians.

The types of history written by the ancient Greeks and Romans influenced profoundly all subsequent historiography down to the 18th century. In order to interpret sympathetically this classical historiography, it is necessary to bear in mind the literary conventions that governed this branch of literature. The ancient Greeks distinguished between history and biography. The origin of both forms can be traced back to at least the 5th century BC, and the differences between them were observed throughout antiquity. The writer of history was supposed to aim at giving a true story, but the biographer was entitled to treat historical personages in a manner that resembled legend. There existed, of course, some exceptions. The lives of the early Roman emperors written by Suetonius in the 2nd century AD, while conforming to the traditional, topical arrangement of biographies, constitute an unusually valuable historical source, especially for Augustus, whose correspondence is repeatedly quoted. Yet another distinction was drawn between history and the study of "antiquities," to use a term employed by Varro (116–27 BC), perhaps the greatest of all the ancient Roman scholars. This distinction was already implicit in Aristotle's contemptuous dismissal of history (in his *Poetics*) as a branch of literature dealing with the particular rather than with things of general significance. The histories he condemned provided chronological narratives of wars and political events. Aristotle and his disciples were engaged in several enterprises that they regarded as something quite different from history. For example, they embarked on the study of the constitutions of all the Greek states. Such work was to be based on systematic inquiries. The student of the "antiquities" tried to use a wider range of evidence than the sources normally consulted by the ancient historians, and he arranged his results systematically by topics.

Ancient biography

In antiquity a writer of history was usually preoccupied at least as much with style as with content. A generation before Aristotle, the rules of rhetoric, as they might be applied to history, were fully elaborated by Isocrates, a teacher of rhetoric at Athens. Cicero tried (especially in his *De oratore,* 55 BC) to familiarize the Romans with these Isocratean precepts. History was to be written in a clear but solemn style, akin to fine oratory. The historian was to introduce all manner of literary embellishments but was also to stress the moral lessons of his story. At its worst this type of historiography could lead to serious misrepresentations of the past. Among the Roman historians, Livy (died AD 17) was an important practitioner of this kind of writing, which was particularly well suited to the patriotic myths that he was trying to immortalize, of a Rome that owed its magnificent destiny to the unique virtues of its citizens and the perfection of its antique institutions. Some outstanding historians, such as Polybius (2nd century BC) and Caesar (died 44 BC), eschewed these rhetorical precepts, but in all the ancient writers an important element of literary artifice was always present. This is one of the reasons why they offend modern standards, which demand absolute accuracy in the presentation of evidence. One of the most striking contrasts is the reluctance of the ancient historians to quote documents. Tacitus might rely heavily on the archives of the Roman Senate, but he never mentions his documentary sources. An inscription discovered at Lyons, France, preserves a speech delivered by the emperor Claudius to the Senate in AD 48, and it is clear that Tacitus utilized another version of the same text. His skill in using it is matched by the freedom with which he adapts it to suit his purpose.

The greatest and the most original achievement of the best Greek historians lay in their clear grasp of the need to distinguish truth from fiction and their conscious preoccupation with the methods of achieving this. This is admirably conveyed in a famous passage of Thucydides.

Methods of Thucydides

> And with reference to the narrative of events, far from permitting myself to derive it from the first source that came to hand, I did not even trust my own impressions, but it rests partly on what I saw myself, partly on what others saw for me, the accuracy of the report being always tried by the most severe and detailed tests possible. My conclusions have cost me some labour from the want of coincidence between accounts of the same occurrences by different eye-witnesses, arising sometimes from deficient memory, sometimes from deficient impartiality.

His practice did not fully live up to this ideal, however. The greatest of his Greek successors, Polybius, is reasonably impartial, except in his treatment of some of the events in Greece. Among the Romans, the writing of history was chiefly the preserve of members of the senatorial class, who almost invariably had some personal axes to grind. But the correctness of the rules formulated by Thucydides was accepted, in principle, by most ancient historians.

Thucydides had deliberately restricted himself to the history of his own time, and many of the subsequent ancient historians did likewise. They could depend on their own experience or could question well-informed contemporaries. The surviving fragments of Livy relating to his own lifetime (64/59 BC–AD 17) are much more vivid and convincing than the earlier books of his history (surviving today only down to 167 BC). The tendency to prefer contemporary history was strengthened by the practical bent of many of these writers. Several ancient historians were men of action familiar with warfare and politics. Interested in history as a source of instruction for statesmen, they could write with authority only about wars and political transactions of their own time. Polybius, the exiled Achaean general and a great traveller, derides unpractical, sedentary historians such as Timaeus, who had been writing about the peoples of the western Mediterranean without stirring for 50 years from Athens.

The historians of antiquity were much less skillful in dealing with noncontemporary history, for which they relied on older historians. Where none was to be found, they felt lost, as Livy complains in the early portions of his Roman history. The modern recourse to non-narrative sources was alien to the habits of most ancient historians. They were usually incapable of doing this successfully, just as they were ill equipped to discuss critically the sources used by the older writers.

Herodotus chose for his theme the successful resistance of the Greeks against the Persians at the beginning of the 5th century BC. Thucydides wrote about the Peloponnesian War, in which virtually all the Greek states became involved in the last decades of that century. These were limited subjects of obvious importance for which it was possible to find ample evidence. The strength of the ancient historians lay precisely in imposing an interesting pattern on the events of a selected period, usually contemporary or fairly recent, for which they had manageable sources. The best of them could thereby achieve a sense of dramatic unity and produce literary masterpieces. The speeches that Thucydides invented for some of the main protagonists in his story are artistically the most satisfying parts of his work, and at times they even seem to recapture the spirit of what might have been said on these occasions. In a superb writer like Tacitus, whose political career had included long periods of frustration and insecurity, one does not look for impartiality or for scrupulous truthfulness but, rather, for fascinating insights into what the development of Roman imperial power from Augustus to Domitian (the period AD 14–96) meant to the proud, sophisticated Roman aristocracy for whom he was writing.

Classical study of "antiquities"

The study of "antiquities," as opposed to narrative history, did not normally produce works of literary merit, and this is probably the main reason why most of them disappeared. One important group of such writings originated with Aristotle and his collaborators, writing in the third quarter of the 4th century BC. They were interested in both literary "antiquities" and in the systematic study of the constitutions of Greek states. They had described 158 different constitutions, though only their account of Athens now survives. A comparison of its two main parts illustrates the contrast between the deficiencies of ancient historiography and the impressive achievements of the antiquarian researchers. In the introductory, historical section, Aristotle was baffled by the problem of dealing with the fairly remote past. For each particular period he tried to follow some contemporary sources. The resultant juxtaposition of several writers differing widely in their political outlook produced an account full of contradictions. The second part, however, containing a systematic description of the Athenian constitution, is a masterpiece of shrewd analysis, as are the empirical portions of Aristotle's *Politics* (Books IV–VI), which are based on a wealth of concrete examples derived from the different Greek states.

Aristotle inspired in the 3rd and 2nd centuries BC a great mass of philological and antiquarian research. The most important scholars were to be found in the new Hellenistic states, especially at Alexandria in Egypt and at Pergamum in Asia Minor. Among the surviving Hellenistic fragments, there are commentaries on Herodotus and Thucydides. The Hellenistic scholars were interested in many subjects connected with history and did pioneering work in chronology, geography, and topography. They were accustomed to using every kind of source and to quoting documents extensively. Their greatest Roman disciple was Varro, who tried to recover all the vestiges of the old Roman society and to make a systematic survey of Roman life based on the evidence provided by language, literature, religion, and ancient customs. Most of his writings have been lost, but he supplied the conjectural (though incorrect) date of 753 BC for the foundation of Rome and knowledge of the probable boundaries between some of the groups whose union produced the city of Rome. Unfortunately, antiquarian researches of such penetrating nature were almost never applied in antiquity to the writing of narrative histories.

Early Christian era. The triumph of Christianity in the Roman Empire during the 4th century assured the predominance of a type of historiography radically different from the works of the pagan Greek and Roman historians. Its origins were Jewish. The Jews were the only people of antiquity who had the supreme religious duty of remembering the past because their traditional histories commemorated the working out of God's plan for his chosen people. By contrast, no Greek ever heard his gods ordering him to remember. It was the duty of every Jew to be familiar with the Jewish sacred writings, which were

Purpose of Jewish histories

ultimately gathered into what became the Old Testament. The writers of these biblical books only gave an authoritative version of what everybody was supposed to know, and they were only concerned with the selection of such facts as seemed relevant in interpreting God's purpose. In addition, the Jews also cherished unwritten traditions. To quote Josephus, a Jewish historian of the 1st century AD, "what had not been written down, was yet entrusted to the collective memory of the people of Israel and especially of its priests."

The Christians took over the Old Testament and added to it an additional body of sacred history. The writers of the four Gospels included in the New Testament were bearing witness to assured truths that the faithful ought to know, and no convincing reconstruction of historical facts is possible from these books of the New Testament. The only avowedly historical book in it is the Acts of the Apostles. The New Testament as a whole represents merely a selection from the early Christian writings. It includes only what conformed to the doctrine of the church when, later on, that doctrine became fixed in one form. Between the Acts of the Apostles, dating probably from the late 1st century, and the writings of Eusebius of Caesarea (died *c.* 340) and his contemporaries in the first quarter of the 4th century, there is an almost complete gap in Christian historiography.

For the Christian writers the story of Jesus, as recorded in the Gospels, represented the fulfillment of the prophecies that could be found in various parts of the Old Testament. The Jewish part of the Bible also assured for Christianity the authority of a long antiquity. The history contained in the two parts of the Bible, now indissolubly linked together, became the only authentic record of God's revelation for mankind, dwarfing into insignificance all the records of other peoples and religious groups. The concept of a universal history had not been wholly unknown to the pagan world, but the Christians were the first to apply it effectively. Christian history had to be a universal history, though of a very peculiar sort, where only one sequence of privileged events, Jewish and Christian, deserved detailed record. The Christian claims must have seemed more extravagant to the pagans than even the Jewish ones. Thus Eusebius stated that the Christians were, in fact, born with the world, anticipating St. Augustine's vision of the city of God existing since the beginning of time.

Universal chronologies

In defending their religion against hostile critics, the early Christians were forced to fit some pagan history into their universal scheme. This was achieved by means of universal chronologies from the creation of the world to each writer's own time. The events of Jewish and Christian history were thus synchronized with the main dates of the pagan myth and history. Sextus Julius Africanus, who wrote in the early 3rd century, is the first Christian writer known to have attempted this feat. He allotted 6,000 years to the whole span of human history and placed the birth of Christ in the year 5500 from the creation of the world. This work provided the model for the more elaborate *Chronographia* (*Chronicle*) of Eusebius. It became the foundation for a long succession of Greek chronographies produced by Byzantine writers. A Latin adaptation by St. Jerome (died 419/420) was immensely influential in western Europe for more than 1,000 years. A modern scholar is filled with mingled admiration and despair at the ingenuity of Eusebius and of his more eminent successors and at the absurdity of many of their conclusions. But they did originate and impose on the world a unified scheme of universal chronology. The dating from the birth of Christ was introduced by Dionysius Exiguus, who wrote at Rome in the early 6th century, and it was successfully popularized in the 8th century by the English historian Bede.

The writing of history of their own time was not an essential task for the Christians of the 4th and 5th centuries. When they did so, they wrote primarily in defense of their religion against the pagan world or against rival Christian groups branded as heretical. All these histories belong to religious apologetics. They suffer from inevitable distortions in the choice of what should be mentioned and what must be suppressed, and they are often excessively unfair to outsiders and opponents. These faults were not uncom-

mon among the classical historians, though the Christians were somewhat unusual in their extreme conviction that they alone must be right. A comparison between the Christian historians and an outstanding pagan writer, such as Ammianus Marcellinus (second half of the 4th century), who was very ready to admire those Christians who merited it, brings out the intolerance and narrowness of outlook of his Christian contemporaries.

Eusebius was the earliest and the most important of the Christian historians of the 4th century. He is quite frank about the practical and apologetic aims of his *Historia ecclesiastica* (written 312–324; *Ecclesiastical History*) designed to show how, through a long series of acts of Divine Providence, a Christian empire was finally brought into existence by Constantine. He admits that "we shall introduce into this history in general only those events which may be useful first to ourselves and afterward, to posterity." This work, like his other historical writings, is a mixture of devout fiction and invaluable detail. But there is plenty of the latter in *Ecclesiastical History.* Contrary to the usual practice of the ancient historians, Eusebius tries to specify his sources, and he quotes from them extensively in order to document as fully as possible the developments that resulted in the triumph of Christianity. He provided in this respect a valuable model for his medieval successors. The most astonishing thing about Eusebius was his capacity to handle his sources critically, in matters where it seemed permissible to do so. In one passage of his *Chronicle* he sets aside the authority of St. Paul in favour of a piece of evidence contained in the Book of Judges. In later patristic literature nothing similar is found.

Biography, as it was habitually written in antiquity, could be readily adapted to Christian purposes. St. Jerome modelled himself on Suetonius in compiling the lives of 135 Christian writers (written in 392) as a way of demonstrating the high level of culture attained by his coreligionists. The ancient biographers had freely mingled fact with fiction for the edification of their readers and could be readily imitated by the writers of the lives of Christian saints. The life of St. Anthony of Egypt by St. Athanasius (mid-4th century) set the pattern for this most popular type of medieval literature.

St. Augustine

St. Augustine, the greatest of the Latin Church Fathers of the 4th and 5th centuries, was certainly not concerned with writing of history in any ordinary sense of the term. In his *De civitate Dei* (*City of God*) he might invoke historical evidence to demonstrate the utter degradation of all the non-Christian societies, and he encouraged his pupil Orosius to develop this theme more fully in the latter's *Historiarum libri VII adversus paganos* (*Seven Books of History Against the Pagans,* to 417). Nearly 200 manuscripts of Orosius have survived, testifying to the immense popularity of his work in the Middle Ages. Augustine's greatest influence on historiography lay in his main message. His vision of the divine and the earthly cities confronting each other dominated the outlook of all the medieval Christian thinkers and profoundly affected their treatment of history. Within that divine plan for the world, purely secular history seemed an insignificant thing.

Early China. The preservation of some records of historical events can be traced in China to at least the early part of the 1st millennium BC. Confucius (551–479 BC) was credited, rightly or wrongly, in the later Chinese tradition with editing the annals of his native state of Lu. But the appearance of the first works fully deserving the name of histories resulted from the unification of China under a single ruler in 221 BC. The first such work to survive, the *Shih chi* ("Historical Records"), dates from *c.* 85 BC. Its author, Ssu-ma Ch'ien, is quite justifiably called the father of Chinese historiography. His history exhibits many of the main features of the later Chinese official histories as they continued to be written down to the deposition of the last Chinese imperial dynasty in 1911. Within this fairly unified tradition, China produced a mass of historical writings unequalled by any other country before modern times. Until the late 19th century, Japanese historiography formed an offshoot of this tradition.

The "Historical Records" of ancient China

Chinese scholars showed an interest in the history of China from the earliest times. According to the Chinese conception, history makes sense only if it can furnish practical directives for action or supply correct information upon which action can wisely be based. All the schools of Chinese thought quoted the lessons of history. Confucius, with his stress on the moral content of these lessons, formed part of this universal belief in the value of history. One of the duties inculcated by him was the scrupulous transmission of authentic records. When, some centuries after his death, the unified Imperial state began to recruit its bureaucracy among the Confucian scholars, the recording of all the necessary information and the careful preservation of records became one of the main functions of the Chinese government, both centrally and locally. A long series of official histories and of records connected with them has survived from the time of the T'ang dynasty (618–907) onward. From then on, the great bulk of Chinese history was written by bureaucrats for bureaucrats. From a practical point of view this immense body of historical writings fulfilled a very useful purpose. Such histories were bound to be highly stereotyped and restricted in content to what interested the higher officialdom. It is easy to condemn it by modern Western standards for its excessive preoccupation with concrete details and inability to produce works of wider synthesis. But this Chinese tradition did gradually evolve in the direction of greater rationality and subtlety. Its scope widened as the sphere of government expanded. Furthermore, within this tradition there appeared from time to time writers of genius, men of bold critical spirit, genuine historical insight, and overriding integrity. One of the greatest was Liu Chih-chi (661–721), the writer of the *Shih t'ung,* the first thorough treatise in Chinese, or any other language, on historical method, which also constituted in effect a history of Chinese historiography. He had a successor in Ssu-ma Kuang (1019–86), the author of the first fairly comprehensive general history of China (covering the years 403 BC–AD 959). In the 17th century a remarkable group of historical scholars virtually founded a school of critical Chinese philology. None of these writers succeeded in radically transforming Chinese historiography, but they created an increasingly sophisticated and critical tradition. Their successors in the 20th century assimilated some valuable features of modern Western historiography.

MEDIEVAL HISTORIOGRAPHY

Europe from the 5th to the 11th century. The period stretching from the 5th to the 11th century was a time of very profound cultural decline in regions that had once constituted the western half of the Roman Empire. Almost all the inhabitants of these provinces again became illiterate. There are long periods for which there are virtually no narrative sources, and the bulk of surviving historical writings consists merely of meagre factual annals. Virtually all the writers were ecclesiastics, in marked contrast to the Byzantine lands, where a strong tradition of lay historiography persisted throughout the Middle Ages. The annalists and chroniclers of the West were predominantly monks, and their lack of experience of the secular world outside their cloisters made them into blinkered and unpractical historians. This was true even of Bede, an Anglo-Saxon monk, who was by far the greatest historian of the early Middle Ages.

All the historians of this period were seriously affected by the cultural decline around them. They were having to write in part for a more uncultured audience. Sulpicius Severus, probably the best Western historian of the early 5th century, still intended his *Chronica* (to 403) for educated Roman Christians, but his life of St. Martin of Tours is a piece of medieval hagiography. This model could inspire lives full of folklore and miracle, from which the real human personalities of the saints were almost wholly absent. The same duality of purpose is a notable feature of Bede's voluminous writings. He explicitly recognized that he must adapt himself to his audience when he explained that he was writing in a simple Latin style so that he might be more easily understood by his Anglo-Saxon readers. There is a marked contrast of tone between his theological and his historical writings. As a theologian, Bede follows Eusebius and the earlier Church Fathers in not exaggerat-

ing the frequency of miracles and in believing that they were most common in the earliest days of Christianity. But Bede's lives of the English saints and his *Historia ecclesiastica gentis Anglorum* (*Ecclesiastical History of the English People*), covering chiefly the years 597–731, are full of miracles and visions. There is one or other on almost every page. It is possible that some of these incidents were included by Bede because he thought that his readers expected mentions of these familiar, traditional stories.

Bede's *Ecclesiastical History*

In preparing his historical works, Bede not only took great care to assemble the widest possible collection of sources but also tells the reader what he is using. In dedicating his *Ecclesiastical History* to King Ceolwulf of Northumbria, he requests that

> in order to remove all occasions of doubt about those things I have written, either in your mind or in the minds of any others who listen to or read this history, I will make it my business to state briefly from what sources I have gained my information.

An impressive list follows, including mentions of documents copied for him by friends at Rome, Canterbury, and other places. Like Eusebius, on whom Bede modelled himself, he quotes some of the documents integrally. Bede's methods of securing and recording information are so similar to the practices of modern historians and the judicious tone of his writing is so impressive that the reader is almost taken in into treating him as if he were a modern scholar. But Bede's *Ecclesiastical History* was written as a work of edification in order to strengthen the faith of his readers in Divine Providence, through which, as he saw it, his Anglo-Saxon countrymen had been converted to Christianity. All matters not connected with his main theme are ignored. Bede's handling of evidence on subjects that he regarded as embarrassing inspires mistrust. But these are small matters in comparison with the enormous mass of information that he alone has preserved and the encouragement that Bede continued to give for many centuries to the writing of history.

The influence of Bede and other Anglo-Saxon scholars was greatly felt during the later 8th and the 9th centuries in the Frankish kingdom, where under Charlemagne and his successor, Louis the Pious, there was a modest revival of historical writing. Besides the annals kept at various monasteries, which tended to convey information in a manner that suited the Frankish rulers, there were a few more ambitious ventures. The important *Historia Langobardorum* (*History of the Lombards*), written *c.* 774–785 by Paulus Diaconus, or Paul the Deacon, was the work of one of the best educated men of the time. Nithard, a grandson of Charlemagne, left an invaluable narrative of the disintegration of the Carolingian state during his lifetime. The work that exerted the greatest influence on the medieval writers of biographies was Einhard's *Vita Karoli Magni* (written *c.* 830–833; *Life of Charlemagne*). The author was a leading official and a close companion of Charles, and his work was naturally intended as a eulogy of the great king. Einhard says that Charlemagne retreated safely from Spain, returning with his army safe and sound, except that on a ridge of the Pyrenees, on the way home, he happened to experience some small effects of Gascon perfidy. Nobody would gather from this that the Franks had narrowly escaped a major disaster. Einhard was merely echoing the story told in the semiofficial contemporary annals. Another source of distortion was Einhard's use of a classical model, the *Lives of the Caesars* by Suetonius. The subject headings under which he described Charles and even the very words used were partly borrowed from the lives of Roman emperors, but his Charlemagne is probably in essentials an authentic and credible portrait.

If bulk alone is to be taken as a criterion, annals were the main product of medieval historiography. The annalist merely sets down the most important events of the current year. In the case of the earliest medieval annals, the events were often noted down in Easter tables, in the blank spaces between the dates calculated for the forthcoming Easters. Such paschal annals would be extremely brief. When, as often happened, annals came to be written down in separate manuscripts, distinct from the Easter tables, there was room for the expansion of individual entries. In either case, the resultant annals cannot be regarded as history since the events are necessarily recorded in isolation. But they preserve in a right order the essential facts, which could be rearranged into a continuous narrative. Such a narrative, if it still followed the chronological arrangement of its various annalistic sources, should properly be termed a chronicle.

Medieval historians show little awareness of the process of historical change. They were unable to imagine that any earlier age was substantially different from their own. The unawareness of the meaning of anachronism helps to explain the strange wanderings of medieval annals and chronicles. If a religious community wanted to acquire a historical narrative, it copied some work that happened to be most readily accessible. A continuation might then be added at the manuscript's new abode, and, later on, this composite version might be copied and further altered by a succession of other writers. Hence there are at least six main versions of the annals known as the *Anglo-Saxon Chronicle*. They all derive from the annals kept down to 892 at Winchester, the West Saxon capital. Thereafter, copies were acquired by religious centres in the most diverse parts of England, and one manuscript was being kept up to date at the abbey of Peterborough as late as 1154. An extreme case of wanderings is represented by the annals of the cathedral church of Cracow, the medieval Polish capital. The first section is based on Orosius, the next comprises annals beginning with the death of Bede and containing notices of Frankish and German events, while the Polish section starts with the conversion of Poland to Christianity (965–966) and ends in the 13th century.

The *Anglo-Saxon Chronicle*

Europe from the 12th to the 14th century. Historians are accustomed to regarding the late 11th and 12th centuries as an age of intensified progress in culture and learning; this development, however, did not greatly affect historiography. There was a modest revival of interest in some of the ancient Latin writers, but would-be historians were unsure which ancient models they ought to imitate. A whole series of attempts was made to apply to other races the theme in Virgil's *Aeneid* of a noble group of people guided by the gods toward a splendid destiny. The first essential step was to establish the descent of one's nation from the ancient Trojans and then to trace subsequent history through a series of heroic conquests. The most ambitious of these writings was the *Historia regum Britanniae* (*History of the Kings of Britain*), by Geoffrey of Monmouth (died 1155), which attempted to establish for the Celts a historical destiny greater than any other. Although some, even contemporary, readers were not deceived by the work, and William of Newburgh, one of the best English historians of the 12th century, denounced it as a tissue of absurdities, many seriously accepted it as history.

With a few exceptions, the ablest minds of the 12th century were attracted into enterprises that ignored history; they were more concerned with systematization of thought and with philosophical speculations. One of the exceptions was Otto, bishop of Freising, in Bavaria. He was a grandson of the Holy Roman emperor Henry IV. He received the best education that his age could give, but he was also briefly a Cistercian monk during the most austere period of that order's history. Otto was torn between conflicting impulses to seek the city of God as the only reality and yet to hope for the progress of the German empire. Out of this conflict came his first work, *Chronica* (*The Two Cities*), a chronicle of world history to 1146, perhaps the most profound medieval attempt at a Christian philosophy of history. As Otto himself confessed, it was composed "in bitterness of spirit . . . in the manner of tragedy." The election in 1152 of his nephew and friend Frederick Barbarossa, as emperor, filled Otto with a new elation. The excellence of his second work, *Gesta Friderici I imperatoris* (*The Deeds of Frederick Barbarossa*), derives in a considerable measure from a quality rare in medieval historians, a sense of optimistic belief in the value of writing history because it might become a record of human progress. *The Deeds of Frederick Barbarossa* contains a penetrating analysis of the problems encountered by the

The works of Otto of Freising

German rulers in trying to rule the precociously urbanized Italian society.

As in antiquity, the best medieval works were accounts of contemporary history by men who had participated in the events that they were describing. It is, however, very significant that some of the writers that are prized most highly today survive in only very few manuscripts and were presumably not appreciated by most of their contemporaries. One such work was the *Historia pontificalis* ("Pontifical History") covering the period 1148–52, of John of Salisbury, one of the most accomplished scholars of his age, who was writing about the period when he was in the papal service. Another instance of undeserved neglect is furnished by the *Liber de regno Siciliae* ("Book of the Kingdom of Sicily") covering the period 1154–69, written by an anonymous member of the Sicilian court.

Unlike the ancient historians, the medieval writers of contemporary history had no inhibitions about extensively quoting official documents. In England, a succession of writers preserved a large quantity of such texts. Roger of Hoveden was, in the last quarter of the 12th century, treated by the English kings as a kind of court historian. He preserved valuable legal and administrative records with which he was familiar through his activities as a royal official and justice. Matthew Paris, the most important English monastic historian of the 13th century, was highly regarded by King Henry III and had excellent sources of information. He left behind a collection of transcripts of royal and ecclesiastical documents that today fills a large printed volume. Some writers made their chronicles into an anthology of official records, thinly connected by the author's brief comments. Such is the chronicle of Robert of Avesbury, consisting mainly of the military dispatches of King Edward III and other interesting documents to 1356. Another variant of the same method was for a wholly mediocre chronicle to incorporate exciting pieces of eyewitness narratives by other writers. A dull English monastic product of the late 14th century, the *Anonimalle Chronicle*, includes a narrative of the Peasants' Revolt of 1381, which is one of the most dramatic and interesting eyewitness accounts to be found in medieval historiography.

The most popular histories of the 13th and 14th centuries were encyclopaedic compilations giving all the important facts neatly arranged under the dates of popes, emperors, and other rulers. There were even more ambitious ventures aiming at summarizing all the important facts from all the different branches of human activity. The Dominican Order, created at the beginning of the 13th century, was especially concerned with producing such aids for the dissemination of useful knowledge. The best known of these Dominican works is the immense *Speculum historiale* ("Mirror of History"), by Vincent of Beauvais, written under the patronage of King Louis IX of France. It is a compilation made up of excerpts from many authors.

The 13th and 14th centuries were not a period of any fundamental innovations in the techniques and nature of historiography, but there was a growing diversity of types of historical writing. Very detailed, chatty narratives multiplied, often badly organized and inaccurate, but conveying the authentic atmosphere of the times and vividly portraying leading personalities. Such were the St. Albans chronicles of Matthew Paris (to 1259), the reminiscences of Joinville about St. Louis during the Seventh Crusade (1248–54), the Lombard chronicle of Fra Salimbene (to 1287), or the vast history of the first part of the Hundred Years' War written in the second half of the 14th century by Froissart. Memoirs and histories written in vernacular languages, such as those of Joinville and Froissart, came to be quite common. Laymen began to write histories. Some were great men, like Geoffroi de Villehardouin, one of the leaders of the Fourth Crusade (which captured Constantinople 1202–04), of which he wrote an account. Important urban chronicles began to appear, such as the Florentine chronicle of Giovanni Villani, with its invaluable statistics of Florentine population and activities around 1338. The extraordinary personality of St. Francis, who died in 1226, inspired lives of him more convincingly human than any previous medieval biographies of saints.

Vernacular histories

The Humanist historians of the 15th century tried to make a deliberate break with the tradition of medieval historiography. By their insistence on a more coherent arrangement of subject matter, by their superior critical outlook, and, above all, by their much more accurate awareness of the process of historical change, they had introduced innovations of fundamental importance. In part they owed their grasp of these new possibilities to the influence of Byzantine scholars. In historiography, as in other matters, the new humanistic scholarship was a joint product of Western and Byzantine traditions.

BYZANTINE HISTORIOGRAPHY

During the millennium that elapsed between the collapse of the Roman Empire in the West in the 5th century AD and the Italian Renaissance of the 15th century, in no part of Europe did the writers of history consistently maintain as high a standard of achievement as in the Byzantine Empire. Parts of the 7th and the 8th centuries form lengthy gaps in the record of Byzantine historiography, but this seems mainly to be the result of subsequent losses of manuscripts. When, in the middle of the 9th century, Photius, future patriarch of Constantinople, compiled a record of some 280 books that he had read, he mentioned works of 33 Greek historians, dating mostly from the late Roman Empire and the Byzantine period, 20 of which are now lost. But, among the Byzantines of the 7th and 8th centuries, there was certainly no parallel to the Dark Ages in western Europe.

Sources of Byzantine historiography

The Byzantine historians were heirs to the combined traditions of classical Greek writing, of the subsequent Hellenistic historiography, and of the Christian historical writing of the 4th century. Few ancient Latin historians were ever translated into Greek, and their influence on the Byzantines was, therefore, very slight. The older classical Greek historians provided the Byzantines with their cherished models of language and style. Like all educated Byzantines, the historians continued for a millennium to write in a literary language that soon became unintelligible to the vast majority of their compatriots. Hence, from the 6th century onward, there appeared, side by side with the learned historiography, a succession of popular chronicles written in the ordinary language. Most of these popular writings form—in their prejudice, ignorance, and crudity—a startling contrast to the works of the more eminent classicizing historians, but they do provide valuable glimpses of the sort of hagiographical history, more religious myth than sober fact, that ordinary Byzantines apparently wanted to read.

Herodotus and Thucydides were frequently invoked by Byzantine historians as models of fine prose. The influence of these two writers on the substance of what was written usually remained slight and superficial, however. The only Byzantine writers who seriously modelled themselves on these two oldest Greek historians wrote during the 15th century. The earlier Byzantine historians owed most to Polybius and to the Greek biographer Plutarch (died *c.* AD 119), the two Hellenistic writers who had the greatest influence on Byzantine notions of how history and historical biography should be written.

Like Polybius, the majority of Byzantine historians, including most of the best ones, perferred to write about their own times; and within these limits they produced some real masterpieces. Unlike the majority of the ancient historians, Polybius had included much autobiographical detail, and his influence reinforced the readiness of the Byzantine historians to talk about themselves, thus providing abundant information about several of these authors. Their histories are likely to be one-sided and full of details about what interested them, while remaining silent about a great mass of other contemporary happenings. They are frequently gossipy and patently prejudiced, inspiring much less confidence than the austere, impartial writings of authors such as Thucydides. This is one of the main reasons why the Byzantine historians have often been excessively underestimated by modern readers. The bulk of the Byzantine contemporary histories were written by statesmen, high officials, and prelates—men with access

to important information. They have to be used critically and cautiously but can be immensely valuable.

Procopius' histories

Priscus of Panium (*c.* 450), a member of a Byzantine embassy to Attila's camp, is the best source of information about that terrible king of the Huns and his followers. A century later, the reconquest of Vandal Africa and of Ostrogothic Italy by the emperor Justinian was the main theme of the *History of the Wars* of Procopius, a leading civilian adviser of Belisarius, the Byzantine commander. Subsequently, Procopius also wrote a *Historia arcana* (*Secret History*), containing a horrible indictment of the activities of Justinian and Belisarius. Many of his details about the corruption at court and the oppressive nature of the government may be substantially correct. In the 11th century Michael Psellus, who wrote a history of his own times, was a leading Byzantine scholar and official, for a time even the chief adviser of emperors. His *Chronographia* is concerned almost entirely with the happenings at the Byzantine court and is one of the most gossipy and amusing narratives ever written on such a subject. His psychological insight and his lively and subtle style delighted the educated Byzantines. Anna Comnena, the daughter and biographer of the emperor Alexius I, greatly admired Psellus. Her own *Alexiad* is a much less fascinating work, but the recovery of the Byzantine power under her father provided her with an important theme.

The last, increasingly disastrous, centuries of Byzantine history are recorded by a series of scholarly and interesting historians. Nicetas Choniates, a high imperial official, provides a surprisingly balanced eyewitness account of the siege and capture of Constantinople by the forces of the Fourth Crusade (1202–04). George Acropolites, a leading adviser of the Greek emperors of Nicaea, carries the story from 1203 to the recapture of Constantinople by the Byzantines in 1261. The later 13th and 14th centuries are covered by a succession of writers deeply immersed in contemporary theological disputations. Perhaps the most readable of all Byzantine histories is the largely autobiographical work of the leading politician and emperor John VI (reigned 1347 to 1354), written after his deposition during his years of enforced retirement in a monastery. George Sphrantzes, a close friend of the last emperor, Constantine XI, included in his history an eyewitness account of the siege and capture of Constantinople by Mehmed II in 1453. Two of Sphrantzes' contemporaries chose to write primarily about the Turks. Their methods place them among Renaissance historians. Laonicos Chalcocondyles wrote (in about 1464) an account of the rise of the Turkish state. He did so in the manner of Herodotus, with long digressions on various neighbouring nations. A little later, Critobulos of Imbros, in his account of the Turkish conquest of Constantinople, made Mehmed II his chief hero and modelled his history on Thucydides.

The study of what might be called "historical antiquities" was not much cultivated by Byzantine scholars. The most notable exception was the emperor Constantine VII, but only some fragments of his voluminous collections have survived (dating from about 940 to 959). They include a very interesting account of the various peoples with whom the Byzantines had to deal. Such ancient Greek literature as still survives, including that of all the historians, was preserved by the Byzantine scholars. When, around the year 1400, the teaching of Greek was introduced into Italian universities by Byzantine scholars, they brought also their superior techniques of literary scholarship, transforming thereby the study of Latin authors as well as introducing into western Europe the treasures of Greek literature. One result was the emergence of the new Renaissance historiography.

MUSLIM HISTORIOGRAPHY

Muslim historiography appears to have originally developed independently of European influences. Until the 19th century Muslim writers only very seldom consulted Christian sources and almost never noted events in Christian countries. Fortunately, they displayed at times more curiosity about the non-Muslim peoples of Asia. The first and best history of the Mongol conquests in the first half of the 13th century was the work of a Persian, Joveyni. On a visit to Mongolia in 1252–53, he was able to consult the recently compiled, earliest Mongol narrative (*Secret History of the Mongols*).

Origins of Arabic historiography

The origins of Arabic historiography still remain obscure because of the gap between the legendary traditions of pre-Islāmic Arabia before the start of the Muslim era (AD 622) and the sophisticated and fairly exact chronicles that began to appear in the later 8th and 9th centuries. But while the detailed stages of this development still await reconstruction, the main influences shaping the early Muslim historiography are clear enough. As in the case of the ancient Jews, it was created and perpetuated by religion. Muḥammad (died 632) regarded himself as a successor to a long series of Jewish and Christian prophets, and he made Islām a religion with a strong sense of history. The Qurʾān, Islām's holy book, is full of warnings derived from the lessons of history.

Teachings of Muḥammad not included in the Qurʾān came to be regarded after his death as authoritative tradition left behind by him. All his sayings and actions were therefore carefully treasured and ultimately came to form, in combination with the Qurʾān, the foundation for the body of Muslim law (Sharīʿah), common to all Islāmic communities. These traditions (Ḥadīth) were transmitted orally for several generations, until they were written down in the 8th and 9th centuries. The resultant collections were only partly historical, as myths and inventions crept into them. The scholars who were engaged in preserving and verifying these traditions were chiefly preoccupied with organizing them into legal and theological systems, and they were frequently hostile to the historians. The earliest authoritative life of Muḥammad, written by Ibn Isḥāq (died 768), was attacked by a leading exponent of the legal "traditionist" learning. This confirms the independence of the historical scholars from the theological and legal interests. But both groups shared some common materials, and the strict rules evolved by the legal "traditionists" for recording their sources and tracing a continuous chain of authoritative transmitters of the traditions encouraged similar exact habits in the Muslim historians. The resultant histories were often pedantic, full of unrelated facts, and deficient in reflective comment, though there are some astonishing exceptions, such as the writings of Ibn Khaldūn (1332–1406). But the better Muslim historians scrupulously quoted their authorities and tried to be truthful. This was particularly true of the "classical" school of historians, who were writing at the centre of the ʿAbbāsid caliphate in Iraq in the 9th and 10th centuries. Aṭ-Ṭabarī (died 923), the most authoritative of them all, wrote his "History of Prophets and Kings" as a supplement to his earlier commentary on the Qurʾān, and subsequent Muslim historians were content to follow his reconstruction of the early Islāmic history. The Syrian and Iraqi historiography of the 12th and early 13th centuries is at least as valuable as the Western historical writing of this period, and sometimes it is clearly better.

To orthodox Muslims, the development of the Islāmic community represented a continuous manifestation of God's purpose. Consequently, the recording of the religious progress of the Islāmic society continued to be sacred duty. One of the original features of Muslim historiography is the large amount of attention devoted to the lives of devout men and of scholars. To many Muslim historians, these spiritual and intellectual activities were of much greater importance than the doings of princes and warriors. One of the peculiarities of Muslim historiography was the liking for encyclopaedic dictionaries of famous men. The earliest of these were devoted to the Companions of Muḥammad and to the early transmitters of the Muslim traditions. For a thousand years extremely diverse types of biographical collections have continued to appear in the Muslim world. Those devoted to religious scholars attained a particularly wide diffusion. Saladin (Ṣalāḥ ad-Dīn), who took Jerusalem from the crusaders in 1187 and later opposed the Third Crusade, offered to the Muslim writers the particularly congenial subject of a ruler dominated by a sense of religious duty. A particularly fine example of medieval Muslim historiography is the biography of Saladin by Bahāʾ ad-Dīn (died 1234), which gives

an exceptional insight into Saladin's motives for many of his critical decisions.

The works of Ibn Khaldūn

But the greatest Arab historian and one of the most penetrating thinkers about historiography in any time or place was undoubtedly Ibn Khaldūn. The introduction (*al-Muqaddimah*) to his *Kitāb al-ʿibar*, a universal history (begun in 1375), is, in A.J. Toynbee's judgment (1934), "the greatest work of its kind that has ever yet been created by any mind." Ibn Khaldūn had absorbed all the learning accessible to a Muslim of his time. He was a master of religious learning, an outstanding judge, a writer on logic. He turned a subtle and most disciplined mind to historiography in order to explain his personal tragedy. He had served a succession of rulers in Islāmic Spain and the Maghrib (Northwest Africa) as a general, a politician, and even once as a chief minister, and his activities had always ended in disaster. In order to explain what had gone wrong, he sought to achieve a correct understanding of the forces that governed the societies known to him. He concluded that political stability had become impossible in his native Maghrib, because over centuries economic prosperity had declined excessively and the forces of lawlessness had become too strong.

As a detailed chronicler of events Ibn Khaldūn is not always exact, but, like contemporary historians, he knew how to reconstruct correctly the main trends over several centuries. His ability to formulate general laws that govern the fate of societies and to establish rules for the criticism of sources provided him with an intelligent framework for the correct reconstruction of past history.

Ibn Khaldūn's *Muqaddimah* has survived in at least a score of manuscripts, but he has had no effective influence on Muslim historiography until recently; after his time, as before, the writing of history continued to be a normal feature of Muslim civilization in the more advanced Islāmic societies. In several countries, notably in parts of India, the first works that deserve the name of history appeared only after the Muslim conquest or the conversion to Islām. After the 12th century Arabic ceased to be the main language of Muslim historiography. Distinguished histories were written in Persian in the 13th century, and subsequently Turkish and other vernaculars came to be used by historians in different parts of the Islāmic world. But, in its isolation from non-Muslim influences and its traditional interests, Islāmic historiography underwent no intrinsic change until the 19th century, when it began to be affected by the impact of modern Western civilization.

HISTORIOGRAPHY IN THE EUROPEAN RENAISSANCE

If there is one thing that united the men of the Renaissance, it was the notion of belonging to a new time. Lorenzo Valla, one of the ablest of the early Humanists, in a preliminary draft of his history of King Ferdinand I of Aragon (written in 1445–46), proudly enumerates the modern technical inventions made in recent centuries, and especially near his own day. The sense of the novelty and excellence of their achievements was particularly felt by the men of the Renaissance in connection with their attempts to imitate the works of the ancient Greek and Roman writers and artists. They were not yet claiming that an era of unlimited progress was dawning for mankind—such concepts belong to the 18th century—but the belief in the progressiveness of their own age soon spurred the best Renaissance scholars and artists into achievements that, in some important respects, surpassed their ancient models. This happened in historiography, and especially in the sciences connected with it. The pace of change must not be exaggerated, however. Despite promising beginnings, historiography as a systematic discipline did not emerge during the Renaissance and, in fact, this development did not occur until the 19th century. The reasons for this delay form one of the main problems in any study of historiography between the years 1400 and 1800.

Denigration of the Middle Ages

In the early Renaissance one by-product of the newly won sense of modernity was the tendency to regard the millennium between the collapse of the Roman Empire in the West and the 15th century as an era of prolonged decline. The concept of the Middle Ages was thus introduced for this intervening period. Two very important histories written in the first half of the 15th century deliberately concentrate on the medieval centuries. Their authors were leading Italian Humanists. The first to appear was the *Historiae Florentini populi* ("History of Florence") of Leonardo Bruni, the city's chancellor from 1427 to 1444. The second, the *Historiarum ab inclinatione Romanorum imperii decades* ("Decades"; mainly devoted to Italy), was written by Flavio Biondo, an important papal official. It covered the period from the sack of Rome by Alaric in AD 410 to the writer's own time. The "invention" of the Middle Ages as a separate historical period remains one of the most enduring legacies of Renaissance historiography.

Unlike the medieval historians, the Renaissance Humanists became much more acutely aware of the process of historical change. This was a gradual development. They were trying to understand the ancient writers, whom they were seeking to emulate, and they became increasingly aware of the need to replace these writers in their correct historical setting. When Petrarch (1304–74), the pioneer Italian Humanist, unearthed in 1345 a collection of Cicero's letters, he was shocked to discover that Cicero was not a cloistered scholar of the medieval tradition but a busy politician who wrote his dialogues in moments of banishment from active life. In 1361, in a letter to the Holy Roman emperor Charles IV, Petrarch was able to use his increased familiarity with classical documents to expose a medieval forgery of the Austrian archduke masquerading as a charter of Julius Caesar.

Between about 1440 and his death in 1457, Valla was one of the most influential Humanists. His *Elegantiae linguae latinae* (1444; "Elegancies of the Latin Language") was a treasury of information about correct Latin usages. For Valla the meaning of words was not natural but conventional and historical, because it was derived from changing custom. Thus a sense of ceaseless historical evolution was planted at the very centre of Humanist preoccupations with the recovery, the correction, and the interpretation of ancient texts.

In 1440 Valla's patron, King Alfonso of Naples, at war with the papacy, asked Valla to write a treatise against Pope Eugenius IV. Valla obliged by decisively disproving, on both linguistic and historical grounds, the genuineness of the "Donation of Constantine." From the middle of the 8th century, when this document was probably concocted, it had been used by the popes as one of the weightiest justifications for their claims to secular authority in Italy. Its authenticity had been sometimes questioned in the past by some of the acutest minds, such as Bishop Otto of Freising in the 12th century and Marsilius of Padua in the first half of the 14th century, but it required Valla's expert techniques to dispose of the "Donation" forever. The validity of Valla's methods of historical criticism was at once recognized by at least one other leading Humanist. Biondo wrote the relevant portions of his "Decades" of papal and Italian history between 1440 and 1443, while remaining in the service of the very same Eugenius IV who had been the chief object of Valla's attack. Yet Biondo tacitly accepted Valla's conclusions, and he never mentions the "Donation of Constantine." Biondo's critical outlook found still another expression in his summary dismissal of the fabulous history of Geoffrey of Monmouth. In his copy of Geoffrey he entered only a single note: "I have never come across anything so stuffed with lies and frivolities."

Historical philology

Valla's work on the texts of the New Testament proved in the long run to be one of the most influential applications of the new science of historical philology. His aim was to recover, so far as possible, the original Greek version through the use of the oldest extant manuscripts. He defended these researches by pointing out that he was not correcting the Holy Scriptures but merely the Latin Vulgate translation of St. Jerome that had been adopted by the Catholic Church. The revolutionary nature of Valla's historical approach comes out most strikingly in his comment that "none of the words of Christ have come to us, for Christ spoke in Hebrew and never wrote down anything." The corrections assembled by Valla became generally known when, in 1505, Erasmus published them as *Annotationes* on the New Testament. They provided a model for Erasmus' edition of a Greek New Testament in

1516, from which stem all the new Protestant versions of the 16th century.

The new historical philology was also soon applied to the study of philosophical and legal texts. In this, the most striking progress was made in the second half of the 15th century by Politian, who lectured at Florence, and by his friend Ermolao Barbaro, who taught at Padua. They were inaugurating the history of ideas and of intellectual movements. In his studies of Aristotelian texts, Barbaro insisted on using only the commentators of antiquity. In his lectures and writings (1489–94), Politian tried to reestablish from internal evidence the correct sequence of Aristotelian treatises, and he traced the gradual liberation of Aristotle's thought from the influence of Plato. The meaning of the terms used by Aristotle was rigorously investigated in the light of the linguistic usage of his Greek contemporaries. Politian's ventures into the field of legal texts proved particularly influential. He had at his disposal a very good 6th-century version of the Digest—that is, the section of Justinian's Corpus Juris Civilis (Body of Civil Law) based on the rulings of the Roman jurists. Politian's collation of it with the first printed edition of the Digest (in 1490) formed part of an inquiry into the transmission of the texts of the Roman law during the Middle Ages. Politian's researches stimulated a remarkable school of Humanist jurists, mostly Frenchmen, headed by Guillaume Budé, who published the first historical commentary on the Digest in 1508. In the course of the 16th century, these scholars laid the foundations of a new branch of scholarship, the history of laws and institutions.

The methods of textual criticism used by Politian and his friends were designed to produce definitive editions of classical texts. Politian was aware of the need to establish the correct descent of manuscripts and to disentangle the best textual tradition. In all this he was far ahead of almost all his contemporaries, and he was anticipating the procedures that were systematically adopted for the first time by Karl Lachmann and other German scholars in the 19th century. The historical philology of Politian was a program for the future rather than a dawn of a new era in the editing of classical texts. In contrast to his methods, most of the other Humanist editions of the Latin and Greek classics are very unsatisfactory. This is particularly true of the editions produced between about 1400 and 1550. The reckless emendations of Humanist editors, coupled with the subsequent disappearance of some of the manuscripts used by them, created grave problems for later scholars. Ever since the 17th century the task of the more modern editors has consisted largely in reconstructing, so far as possible, the manuscript versions available before 1400.

Modern historiography was created in the 19th century through a successful combination of the use of narrative sources with every other type of evidence. Some 15th-century Italian Humanists were already aware of these possibilities. The idea of recovering an entire civilization through a systematic collection of all the relics of the past was not alien to them. Biondo used mainly conventional narrative sources for his "Decades" of Italian history, but his description of the city of Rome in antiquity (*Roma instaurata,* 1444–46) was based on a novel combination of the narratives of other historians with a wide range of miscellaneous sources. These included topographical guides, public and private documents, studies of surviving buildings, inscriptions, and coins. But in practice most histories and biographies continued to be written in a conventional way, while the revived study of "antiquities" was cultivated in separation from narrative historiography.

Biondo's description of ancient Rome

Imitation of ancient models is the feature most often stressed in the modern descriptions of Humanist histories. This meant that style mattered at least as much as content and that historical truth might be obscured by literary conventions. On the more positive side, there was the renewed insistence on the choice of definite, clearly delimited subjects and on a more coherent arrangement of material. The abler Humanist historians, however, were also making innovations that bring their practice a little nearer to present notions of writing history.

Several Humanist historians were particularly attracted to the study of the origins of the states about which they were writing. In the 15th century Bruni did this for Florence, and Biondo and Bernardo Giustiniani for Venice, to mention some notable examples. In the 16th and early 17th centuries, French and English scholars inaugurated a critical study of the origins of their national institutions. Humanist historians prided themselves on their critical ability to overthrow the legends in which various countries had concealed their ignorance of their own origins. The incentives to revise the earliest history were often political. Bruni deemed it essential to prove that Florence had not been founded under the tyranny of the Roman emperors but in the time of the free republic. He happened to be right. The Humanist historians were more confident than their ancient predecessors that they could write competent histories of a remote past. In practice they were much less successful in this than they imagined. In dealing with periods before their own time, they usually followed only a restricted number of earlier narratives, though the best of them, such as Bruni and Biondo, displayed in their histories of medieval Italy a novel ingenuity in combining well-chosen sources. Biondo, for example, made effective use of Dante's correspondence.

There was also some modest progress through the better use of documentary sources. This is often far from obvious, because Humanist historians, like their ancient predecessors, do not usually refer to their sources, even when they quote texts verbatim. Hence came Leopold von Ranke's utter misjudgment of the historical value of the *Storia d'Italia* ("History of Italy") of Francesco Guicciardini. Before Ranke's time it was universally accepted as the most authoritative contemporary history of Italy in the years 1494 to 1534. Ranke, who became one of the pioneers of "scientific" history in Germany, first established his reputation in 1824 by his attack on the reliability of Guicciardini. Ranke argued that the statements of that great Florentine statesman were contradicted by documentary evidence and that his history must have been based on unreliable secondary authorities. The discovery in the 20th century of Guicciardini's private archive proved that his history was scrupulously based on original documents of the highest value.

Guicciardini's history of Italy

Guicciardini, in a work that forms the nearest Renaissance parallel to the history of Thucydides, tries to comprehend the succession of tragedies that befell Italy from the start of the French invasions in 1494. This desire to recapture the rational causes of events is one of the most mature features of the best Renaissance historiography.

EARLY MODERN HISTORIOGRAPHY

Italian Humanist historians provided models that could be imitated easily in other countries. Almost everywhere in western and central Europe, local writers were encouraged to produce descriptions and histories of their own lands, intent with patriotic pride. In such countries as Spain and Poland, which had only recently achieved their unity, this was a way of commemorating their newly won cohesion. In the 15th century it was the object of a pioneer work on the earliest antiquities of Spain, the *Paralipomena Hispaniae,* by the Catalan Humanist bishop Joan Margarit i Pau, and of the invaluable *Annales seu cronicae incliti regni Poloniae* ("History of Poland"), by Jan Długosz, which included an exceptionally precise geographic description of his country. In Germany a sense of national identity could be vindicated by Humanist historians striving to minimize the importance of the continued political division of their land. The *Germania* of Tacitus was printed in Germany as early as 1473 and started the fashion of using this collective name for that country. Tacitus called the Germans "the indigenous inhabitants." This was used by a leading patriotic Humanist, Conradus Celtis, as a proof that Germany should be free from all foreign domination. Celtis and his other Humanist contemporaries deliberately hunted for manuscripts of medieval German writers to prove that their country, despite its disunity, could have a national history. Some important masterpieces were recovered, including the histories of Otto of Freising. Celtis' pet project of a description of Germany modelled on Biondo's *Italia illustrata* was carried out in 1530 by Sebastian Münster, and Münster's

fuller *Cosmographia* (1544; "Cosmography"), though purporting to describe the known world, devoted one-half of its 818 pages to "the German nation." There was also a spate of histories of Germany, mostly very laborious and unreflective but incorporating the newly rediscovered medieval narratives and even some documentary sources. Greater originality came only in the wake of the Reformation. The same thing happened in France and in England. In both countries patriotic preoccupations were a leading feature of works written by Humanist historians, and the appearance of Protestantism reinforced in a peculiar way the existing nationalist tendencies.

The influence of the Reformation on historiography must first be discussed at a more universal level. As the philosopher Francis Bacon shrewdly observed, Martin Luther had been obliged "to awake all antiquity and to call former times to his succours . . . so that the ancient authors . . . which had a long time slept in libraries, began generally to be read." This was not because Luther would have regarded himself as a historian. But as early as 1519, in his disputation with Johann Eck, he encountered the assertion that the primacy of the pope was of divine origin. In order to disprove this and to demonstrate that they alone represented the true church, the Protestants had to retell in a new way the entire history of Christianity. In a preface to the *Vitae Romanorum pontificum* ("Lives of the Pontiffs"), published by Robert Barnes in 1535, Luther himself confessed that, although he himself had not originally attacked the papacy with historical arguments,

> now it is a wonderful delight to me to find that others are doing the same thing . . . from history—and it gives me the greatest joy . . . to see . . . that history and Scripture entirely coincide in this respect.

Protestant history

The starting point for the Protestant rewriting of Christian history could best be found in St. Augustine's teachings. The true church, the city of God, had always existed, even though at times it seemed to be overshadowed by the enemies of the divine order. Those enemies were not only the pagans and the heretics, as St. Augustine had believed. In more recent times they had included also the upholders of the papal authority and the persecutors of such medieval true Christians as John Wycliffe (died 1384) and John Hus (died 1415). The writings of Eusebius provided the model for chronicling the sufferings of the faithful until the dawn of freedom for the true church in the 16th century. These views about the correct history of Christianity were presented with exceptional cogency in John Calvin's *Christianiae religionis institutio* (fullest edition 1559; *Institutes of the Christian Religion*) and were shared by most Protestant scholars. The only obvious disagreements arose when Protestants tried to pinpoint the moment at which the church took the fatal turn away from God's true purpose. While the radical sectarians considered that the papacy had always been corrupt, less extremist Protestants were prepared to accept the earlier popes and to argue that the rot set in at some date between the time of Eusebius (died *c.* 340) and the 7th century. The choice of precise date might depend on the national traditions of each country. Thus, Bishop Richard Davies, in his preface to the New Testament in Welsh (1567), treats Pope Gregory the Great (died 604) as a special enemy because Gregory's effort to convert the Anglo-Saxons led ultimately to the subjugation of the autonomous British church.

Historians writing in this spirit were incapable of impartiality. But the historical controversies between the Catholics and the Protestants produced from both sides huge compilations. Their authors were determined to prove their respective cases by a stupendous marshalling of authorities and documentary sources. The habit of giving copious references and long, exact quotations, missing from the Humanist historiography, was reintroduced by the religious controversialists. On the Protestant side, the largest work is the *Ecclesiastica historia,* or the so-called *Centuriae Magdeburgenses* (13 volumes, 1559–74; "Magdeburg Centuries"), retelling the history of the church down to 1200. The Catholic reply, equally huge and graceless, was produced in 12 volumes by Cardinal Baronius. The chief Protestant critic of this work, the great Greek scholar Isaac Casaubon, was astonished by the Cardinal's ignorance of Greek and Hebrew, his gross mistakes, and his boundless credulity.

The narratives of contemporary events written in the 16th and early 17th centuries by the participants in the religious struggles, though equally partisan, include some works of great historical value and high literary merit. The earliest and best German Protestant narrative, that by Johannes Sleidanus, received a grudging tribute from his great opponent, the Holy Roman emperor Charles V, who remarked that "the rogue has certainly known much . . . ; he has either been in our privy council or our Councilors have been traitors." John Foxe's *Book of Martyrs* (1563) contains a great mass of exact information about the persecution of reformed religion in England and Wales during the reign of Mary Tudor, and it has influenced many generations of British Protestants. The achievements of Queen Elizabeth I and the Anglican Church's settlement of her reign found an outstanding defender in William Camden, who was encouraged to write by Elizabeth's leading ministers. In his *Annales Rerum Anglicarum, et Hibernicarum Regnante Elizabetha* ("Annals of Elizabeth's Reign") Camden made excellent use of a mass of official records at his disposal, though his treatment of confidential matters had to be discreet.

Foxe's *Book of Martyrs*

Out of a conflict between Venice and the papacy in the first years of the 17th century was born the *Istoria del concilio tridentino* (1619; *History of the Council of Trent*, 1676) of Fra Paolo Sarpi. A Catholic friar, but a passionate defender of Venetian autonomy, Sarpi drew a dark picture of worldly papal policies and the unscrupulous machinations of the Jesuits. It is a bitter, prejudiced, but splendidly written and well-informed work, which profoundly influenced the anticlerical historians of the 18th century. All these contemporary narratives, however, have one serious limitation. They deal almost exclusively with political events and with changes in ecclesiastical organization. The Protestant schism is treated as merely a revolt against the abuses of the old church, and the deeper reasons for the alienation of the Protestants from the Catholic faith are never explained. Furthermore, these historians, by attributing the origins of the schism almost exclusively to Luther's sudden conflict with the papacy, obscured the existence in the early 16th century of numerous Catholic reformers, whose sole aim was to transform the Catholic Church from within. This one-sided approach to the history of the Reformation was destined to persist for a long time. Two influential histories published in the years 1683–88, one by a great Catholic prelate, Bishop Jacques-Bénigne Bossuet, and the other by Pierre Jurieu, a leading Protestant, still agreed on the same superficial account of the causes of the Reformation.

The rewriting by the Protestants of universal church history naturally involved a drastic revision of the history of the national churches. In Germany, particularly, the history of the church had become inextricably intermixed with the destinies of the German empire. Their hatred of the papacy made the Lutherans visualize the course of German history with unusual clarity. Nobody before them had attempted to impose on that history a single intelligible pattern of any sort. Theirs was bound to be a prejudiced pattern, a story of gradual national disintegration as the result of the successive defeats of the German emperors by the papacy. Johannes Stumpf's tragic chronicle of the Holy Roman emperor Henry IV (published in 1556) treated his struggles with Pope Gregory VII as the beginning of the empire's tribulations. The whole course of German history was retraced in this fashion under the influence of Luther's chief Humanist collaborator, Philipp Melanchthon, in the so-called *Chronicle* of Carion, written in its final versions (1572–73) by Melanchthon's son-in-law, Caspar Peucer.

One of the most novel features of the English Protestant historiography was the reawakening of scholarly interest in the period before the Norman Conquest of England in the 11th century. Matthew Parker, Queen Elizabeth's first archbishop of Canterbury, thought he could discern in the pre-Conquest church elements of true Christianity that were destroyed thereafter and had only been reintroduced by the Protestants. The Anglican Church could be repre-

sented as a return to the traditional practices and beliefs of the early English Christians. Thus the replacement of Latin by English in the Protestant church services could be justified by citing the presence in Anglo-Saxon England of Bibles, liturgies, and devotional literature in the Old English language. Parker and his friend Lord Burghley, Elizabeth's most trusted minister, gathered around them a circle of enthusiastic scholars, whose work preserved most of the important Anglo-Saxon texts as well as of some leading post-Conquest chronicles. Parker's own method of editing texts horrifies modern scholars, but some of the antiquarian works published by members of this group were of high quality. Camden's *Britannia* (first edition 1586, later much enlarged) was a pioneer work on the topography of Roman and early medieval Britain. The edition by Sir Henry Spelman of the records of the pre-Conquest church councils was the first serious attempt to apply to an important type of early sources the best methods of continental scholarship.

Camden's *Britannia*

The growth of a historical outlook can be traced in the 16th century in many diverse fields of learning. For the first time men were realizing that there was a historical side to every branch of knowledge concerned with human affairs. "I have become aware that law books are the products of history," wrote the French legal historian François Baudouin in 1561. In each branch of study there developed a special historical technique particularly appropriate to it. The most sophisticated scholarship was to be found in the field of classical studies. A group of scholars active in the second half of the 16th century were achieving results much superior to the work of the earlier Renaissance classicists. They combined philological expertise with a determination to reach a really adequate understanding of the ancient Greek and Roman civilizations. A few were Italians, such as Carlo Sigonio, but most of the important works were written in France and in the Protestant centres of Switzerland and Holland. As textual critics these scholars were reacting sharply to the earlier, more haphazard, methods of emending and editing classical authors. They were trying to bring the text of one writer after another to a state of near perfection. Some leading ancient historians, such as Tacitus, benefitted greatly from this treatment (edition of Lipsius in 1575). Though their methods do not quite reach the standards of modern scholarship, they anticipate intelligently many of the procedures more systematically adopted in the 19th century. Isaac Casaubon was the first to point out in his edition of Suetonius (1595) that Einhard's 9th-century life of Charlemagne was modelled on the work of that Roman historian. Casaubon's friend Joseph Scaliger renewed the science of classical chronology (1583) and was the first to reconstruct the original Greek *Chronicle* of Eusebius lying behind St. Jerome's Latin translation. Sigonio's pioneer work on the rights and duties of Roman citizens (1560) was later much used by Theodor Mommsen, one of the founders in the 19th century of the modern study of Roman history.

In the course of the 16th century, non-narrative historical work of the highest originality and complexity was being carried on in the legal faculties of French universities. One important stimulus was provided by the existence in France of different legal systems—the uncodified provincial customs in the north and the written law in the south. The latter ultimately derived from the Roman law, and, in the southern French universities, there arose an eager demand for the introduction of the new Italian methods of interpreting the Roman legal texts. Andrea Alciato, a pioneer in the historical treatment of the Roman law, taught at Bourges from 1529 to 1533, and his pupils founded the "Romanist" school of French legal historians.

Legal histories

Important advances were made in the study both of the Roman law and of the origins of the French legal customs, laying virtually the foundations of a new branch of scholarship, the history of law and institutions. François Baudouin published in 1545 the first historical survey of the development of the Roman legal science. The treatise on the custom of Paris by Charles Dumoulin (published 1539–58) resulted from his advocacy of the codification of the northern French legal customs. It was the first scholarly exposition of a body of customary French law derived from feudal practices, and it amounted to a first comprehensive history of European feudalism. It prompted a series of controversial works by a succession of scholars. The Roman, the Germanic, and the Celtic roots of feudalism all found advocates, and the respective claims of Lombard and Frankish texts to provide the best clues were vigorously canvassed. The complexity of the problems presented by the unravelling of the origins of feudalism dawned on scholars for the first time. The most valuable of these attempts to rediscover the "ancient French constitution" were the researches on "the antiquities of France" of Étienne Pasquier (published 1560–1607), which form a basis for all later study of medieval French institutions.

One of the novel features of European civilization in the later 16th and 17th centuries was a secularization of mental interests. Secular learning could now produce ideas more fascinating to intelligent men than theology. History was one of the most popular types of literature sought by a growing reading public. Several treatises on the proper way of writing history appeared in the third quarter of the 16th century. An anthology consisting of 12 such works, including the famous *Methodus* of the French political philosopher Jean Bodin, was published at Basel in 1576. Nearly 100 years later a "Catalogue of the Most Vendible Books in England" (1657) showed that history books constituted a large proportion of the total works published. It has been estimated that between 1460 and 1700 at least 2,500,000 copies of 17 leading ancient historians were published in Europe.

The late 16th century and the 17th witnessed the publication of several great collections of historical materials. The men who undertook these gigantic tasks often were antiquarians accumulating miscellaneous records rather than historians, but they were supplying materials for generations of future historians. Some of the most important publications of sources appeared in France and the Netherlands. Pierre Pithou was a pioneer in editing materials for the history of the Frankish period. The collections of André Duchesne are a vast storehouse of chronicles and other sources for the study of medieval French history. Le Nain de Tillemont edited 20 volumes of records devoted to Roman and church history during the first six centuries of the Christian Era, which a century later furnished one of the principal sources for Edward Gibbon's work *The History of the Decline and Fall of the Roman Empire*. In 1629 a Belgian Jesuit, Jean Bolland, embarked systematically on the editing of records connected with all the saints whose feasts had at any time been celebrated by the church, and this series of publications has been continued to the present day. In the second half of the 17th century, the French Benedictine congregation of Saint-Maur started an immense series of publications commemorating the history of the Benedictines and of other monastic orders. The greatest Maurist scholar, Jean Mabillon, was accepted throughout Europe as the most erudite historian of his time.

In spite of its popularity among an expanding reading public and of the large number of learned editions of materials that it inspired, history was not, for most of the 17th century, one of the sciences that made men proud of living in a modern age. Immense progress was taking place in mathematics, astronomy, and physics. History not only did not seem capable of much further development, but scientifically minded men were beginning to dismiss it as a branch of knowledge that would never be worthy of serious respect. Mabillon's *De Re Diplomatica* (1681) helped to challenge this pessimistic view, but a further century elapsed before history began to be accepted as an authoritative discipline.

Hostility of rulers toward historiography

One major obstacle to the progress of historiography was the hostility of rulers to publications that did not favour their governments. The growth of an influential reading public made rulers increasingly suspicious of historical writings; for example, the censorship exercised by Cosimo I de' Medici, ruler of Florence from 1537 to 1574, precipitated the decline of Florentine historiography. Comparisons with the past also could be invidious. In 1599 Elizabeth I of England censured an author for describing

the deposition of one of her predecessors, Richard II, 200 years earlier. Fear of possible trouble made highly intelligent scholars into one-sided historians. The great jurist Hugo Grotius avoided in his history of the wars of the Dutch against Spain discussions of the religious aspects. Samuel Pufendorf, the historian of the Swedish conquests, carefully left out the internal developments in 17th-century Sweden.

The scholars who in that century were responsible for the great advances in the mathematical sciences were convinced that their achievements would ultimately give mankind a novel mastery over its natural environment. This is particularly true of Francis Bacon and of René Descartes. Their optimism was laying the foundations for a belief in a possibility of continuous progress without which the purposeful and assured historiography of the 19th century would be inconceivable. But the attitude toward history of most of the leading thinkers and scientists of the 17th century was not helpful to its immediate development. Bacon, who wrote a readable and rationally argued biography of King Henry VII of England, attached no importance to accuracy; for example, he antedated Henry's death by a whole year and could not be bothered to undertake any detailed research. Gottfried Wilhelm Leibniz was a great mathematician, but his attempts to apply science to historiography led to mechanistic constructions from which real human beings were largely missing. Numerous influential thinkers were decidedly hostile to history. Descartes, the most eminent of the anti-historical scientists, was not simply disgusted by the unsystematic and imprecise methods of the historians of his time but also doubted whether, strictly speaking, history could be regarded as a branch of knowledge at all. But it is important to remember that much of the 17th-century criticism of history was an attitude of men who simply had other priorities and were concerned to attack doctrines that, for one reason or another, historians seemed to support. In the late 17th century the most successful defenders of history were the members of certain particularly scholarly Catholic orders. Catholicism rested its authority on tradition to a much greater extent than did its Protestant opponents. For Catholic scholars such as Mabillon, the defense of history became really a defense of their religion. They were trying to show that historians were capable of discovering scientifically demonstrable truths. The decisive publication was Mabillon's *De Re Diplomatica* of 1681. A member of a rival order, the Jesuit Daniel van Papebroch, had challenged (in 1675) the authenticity of the oldest charters of two French Benedictine monasteries, Saint-Denis and Corbie. Mabillon applied his powerful critical intelligence not only to vindicating these documents but also to formulating the general rules that must be used to prove the authenticity of medieval records. He illustrated his rules by admirable examples and stated his conclusions with a candor and a common sense that convinced most readers. Mabillon's survey of the tests that must be applied by scholars covered the writing materials, the scripts (thus founding the science of medieval Latin paleography), the seals and other devices of authentication, the official formulas, and the vocabulary used at different periods. Above all, he stressed that the authenticity of a document usually rested not just on isolated details but on consistent correctness of all its features.

The rules of diplomatics

Mabillon was not just a "historical scientist." He had a passionate interest in the past and a vivid historical imagination. He displayed these qualities abundantly in his last and most important work, the *Annales Ordinis s. Benedicti* ("Annals of the Benedictine Order," to 1066). In the *Traité des études monastiques*, (1691; "Manual of Monastic Studies"), he defended the importance of scholarly work as the principal activity of an elite of Benedictine monks. But it would be an anachronism to regard Mabillon and his chief associates as fully comparable to modern historians. They were constrained by the limitations of their time and of their special position as monks. For example, Bernard de Montfaucon, Mabillon's most important successor, is the creator of the science of medieval Greek paleography. But he shares with most of his contemporaries a complete inability to treat the Old Testament as a historical source.

Historical and antiquarian studies developed in 17th-century England in several very distinctive ways. The political struggles and religious controversies of that period made some issues of older English history into matters of immediate practical importance. The other distinctive feature was the delay in the absorption of European continental learning, so that the great progress made in the study of feudal origins in the 16th century began to affect the thinking of English scholars only by about 1625. But there persisted also elements of continuity growing out of earlier Tudor scholarship. The interest in the Anglo-Saxon church and civilization continued to stimulate important editions of records throughout the 17th and early 18th centuries, including, especially, Sir Henry Spelman's edition of the records of church councils and Sir William Dugdale's *Monasticon Anglicanum* (1655–73), which is still valuable today. Another element of continuity with the Tudor period was the perennial interest of the English notables in heraldry, genealogy, and the antiquities of their native regions. Dugdale's *Antiquities of Warwickshire* (1656) set a pattern and a standard for county histories.

Students of English law and institutions, lacking the stimulus that was provided for French lawyers by the diversity of legal systems and by the notable progress in the study of Roman law in that country, continued to ascribe immemorial origins to the common law of England and to approach the development of English institutions in a completely unhistorical spirit. Among the parliamentary opposition to the Stuarts, these attitudes were part of a belief in the "ancient constitution," which these sovereigns were supposed to be defying. Spelman, who was a devout Anglican and a royalist, though a moderate one, was perhaps the first major scholar to break away from this myth. Under the influence of continental publications and correspondents, he accepted that feudal tenure had been introduced into England after the Norman Conquest and that all the English institutions after 1066 must be redefined in feudal terms. But his discoveries were hidden in a dictionary of antiquarian words (*Archaeologus*, vol. 1, 1626; 2 vol. 1664) and made very little impact until some 50 years had elapsed. Spelman had an acute sense of historical development, and he sadly castigated his countrymen for their lack of it in their attitude to parliamentary origins:

Achievement of Sir Henry Spelman

> when States are departed from their original Constitution and that original by tract of time worn out of memory; the succeeding Ages viewing what is past by the present, conceive the former to have been like to that they live in. (*Of Parliaments*, written in about 1640, published 1698.)

His greatest contribution to English history was to grasp that parliaments had developed out of feudal assemblies convoked by the Norman kings and that the Commons were introduced into parliaments subsequently, as a result of the growing prosperity of the lesser landholders. These views first became generally accessible in the 1664 edition of Spelman's dictionary. They were adopted by Robert Brady (in 1681) and by other partisans of the Stuarts and expanded into a Royalist statement of the English past. Violently polemical though this view was, it did at least lay to rest the myth of the immemorial "ancient constitution." The Whig triumph at the Glorious Revolution of 1688, which established a doctrine that the king ruled by parliamentary consent, led to the neglect of these discoveries for much of the 18th century. This was the common fate of much of the research of 17th-century antiquarians, who were very much ahead of their time and were writing for a limited audience. John Aubrey's pioneer description in the 1670s of the prehistoric sites of Avebury and Stonehenge had to wait two centuries for full publication. Even the best of these antiquarians, such as Spelman and Dugdale, were less critical in their handling of the original sources than Mabillon was. Higher standards were reached by a few of their successors in the early 18th century, especially by Thomas Madox, whose *Formulare Anglicanum* (1702) imitated Mabillon by attempting a systematic introduction to English medieval documents. But this did not save Madox from prolonged oblivion. After about 1730 this English tradition of antiquarian scholarship largely ended and remained unfashionable for most of the 18th century.

HISTORIOGRAPHY IN THE AGE OF THE ENLIGHTENMENT

The impulse given to historiography by the Italian Humanists and the religious controversialists had largely spent itself by about 1715. Men knew again how to write rationally satisfying contemporary histories, though often it needed courage to do so. Much less progress had been achieved in reconstructing the more distant past. Impressive collections of historical materials were being accumulated, but most scholars still lacked the capacity to rethink the thoughts of past generations and thus really to understand them. Mabillon could write with insight about early Benedictine history, as he possessed both sympathy with the subject and adequate technical expertise, but he was exceptional. Spelman had grasped that a particular society would be molded in a peculiar way by its institutions. He could not reconstruct and explain the gradual changes from one set of institutions to a later one, but he was aware of the problem.

Judged by the quality of its historical output, the 18th century was not, on the whole, an age of successful historians, but some of the defects of earlier historiography were beginning to be overcome. There were also losses, however, for some of the achievements of the preceding period were in danger of being forgotten. In the leading countries of western Europe, religious controversies were becoming less important, and a massive secularization of interests took place, which affected even ecclesiastical scholars. The French Maurists continued until 1790 to publish imposing historical collections, but their choice of subjects was determined much less than in the time of Mabillon by religious priorities. The greatest Italian ecclesiastical disciple of Mabillon was Ludovico Antonio Muratori, a social reformer. In a divided country like Italy, the best way of expressing his patriotism lay in reminding Italians of the former greatness of their country. Muratori spent much of his long life on his editions of Italian medieval sources.

The nationalist motivation shown by Muratori was peculiar to Italy and also to parts of Germany, another divided country. Elsewhere in Europe there was a danger that, as men lost interest in constitutional or religious disputes that might be settled by appeals to the past, they might turn away altogether from history or at least neglect long stretches of it. This did happen to some extent in the 18th century. Some of the radical French reformers, such as Jean Le Rond d'Alembert, one of the main inspirers in the 1750s of the French *Encyclopédie,* wanted to jettison completely much of the past. The Marquis de Condorcet, an early prophet of the doctrine of endless progress of mankind and a pioneer historian of European civilization, was a prominent member of a French parliamentary commission that in 1792–93 deliberately destroyed some of the royal records as comprising relics of past servitude.

During much of the 18th century it was safer and easier to publish controversial works of history than it had been in the past. The point is important, as without this greater freedom, the peculiarly radical "philosophical" historiography, so typical of that century, would have been inconceivable. In Italy such writing was still dangerous. Pietro Giannone, the author of an anticlerical history of Naples (1723), was tracked down by the Inquisition and spent 12 years in prison, where he died in 1748. Even the great Muratori, who tried to help Giannone, came into danger of having some of his works banned and had to be rescued by the personal intervention of Pope Benedict XIV. In France, Louis XIV in 1714 imprisoned Nicolas Fréret in the Bastille for alleging (correctly) that the Franks were originally a confederacy of German tribes and not descendants of more illustrious ancestors. Under the successors of Louis, nothing quite so absurd happened again, but critics of the government or the church were often in trouble. Great Britain, Holland, Switzerland, and parts of Germany, on the other hand, provided safe oases where most things could be published. It was no accident that the most independent and historically minded group of German professors should have congregated at the University of Göttingen, founded in 1734, in the Hanoverian territory of the kings of Great Britain.

"Philosophical" historiography

A real renewal of historiography in the 18th century could only come if fresh reasons were discovered for making it again worthwhile. Nationalism could supply one such motive; but this only became decisively influential in the 19th century. An alternative was a historiography inspired by the progress in the natural sciences and based on formulating the general rules governing the development of human societies. The chief features of this "new" historiography were a sense of the unity of all human history, including an interest in the continents outside Europe; a capacity for bold generalizations about the salient features of particular periods or societies; and a preference for topics connected with the progress of human civilization. Condorcet's historical sketch of the progress of the human mind, written in 1794, subdivided all known history into nine periods, each starting with some great invention or with geographical discoveries.

The shortcomings of this "rationalistic" historiography have been rehearsed often enough. For many of its writers it was primarily a weapon of propaganda against their enemies in church and state. Their redeeming virtue was the fearlessly critical attitude to all existing authorities, however august or sacred. The vast scale of their generalizations often precluded any detailed research. This was particularly true of the attempts to write histories of civilization, as the existing collections of printed materials did not cater for such interests, while systematic research in archives was seldom possible in the 18th century. In preparing his pioneer essay on the history of civilization, covering the millennium from the Carolingians to Louis XIV (*Essai sur les moeurs et l'esprit des nations,* 1745–53), the French author Voltaire had to collect bits and pieces from most diverse sources.

One of the most valuable achievements of the thinkers of the 18th century was their capacity to study particular societies as coherent units and to formulate the theory that the various aspects of each society's life were closely interrelated. This was not an entirely novel idea, but it first became commonly accepted during this period. Nor were all its adherents anticlericals. Giambattista Vico, a Neapolitan Catholic, was ahead of his contemporaries in his particularly subtle sense of the complex influences by which one phase of society gives place to another. In his reconstruction of these transitions during the early stages of Roman history, he makes no clear lines between periods. His countryman Giannone explains in his autobiography that he had studied Roman law not for its own sake but in order to understand the changes in the society of the Roman Empire. The French philosopher Montesquieu, who owed much to Giannone, was not really a historian, but he displays an acute sense of historical realities. His *De l'esprit des lois* (1748; *The Spirit of Laws*), more than any other book, accustomed his contemporaries to ponder the complex factors that shaped each society. It inspired Gibbon's definition of the kind of history he wanted to write. It was to be a "history related to and explained by the social institutions in which it is contained."

Gibbon's *Decline and Fall*

This ideal was realized in Gibbon's *History of the Decline and Fall of the Roman Empire* (1776–88), one of the masterpieces of "philosophical" historiography. Gibbon was preoccupied above all with the problem of human progress. The belief that continuous progress was possible for mankind had been publicly formulated in the mid-18th century by Anne-Robert-Jacques Turgot in France and by Adam Smith in Scotland, independently, it seems, of each other. Gibbon had read works and known scholars influenced by both these thinkers. A belief in continuous progress would confer a new purposefulness on the study of the entire course of human history and could justify a lengthy account of what otherwise might have seemed very obscure stretches of the past. Such a justification was to inspire most of the historiography of the 19th century. But the problem of progress had a special urgency for Gibbon's generation, which worried at the thought that their own enlightened civilization might also subsequently collapse. By unravelling the causes of the decline of the Roman Empire, Gibbon was determined to show that the Europe of his own day had attained a much superior degree of development and was immune from the fate of the ancient world.

In the 18th century, historiography was still only very

rarely connected with the universities; and thus, except in such isolated places as Göttingen in Germany, no continuous schools of history could develop. Some of the most important achievements of the 18th-century historians meant much less to their contemporaries than to their successors in the 19th century. Gibbon was a pioneer in utilizing in a "rationalist" history the vast materials accumulated by generations of erudite antiquarians, but he had no immediate followers. The German archaeologist Johann Joachim Winckelmann tried to revive the true understanding of Greek sculpture and to make the history of art into something more than just the biographies of artists, but his work bore little fruit until the next century. The saddest fate was that of Vico's work. He was hardly ever read before the 19th century, when he at last influenced Barthold Georg Niebuhr and the rest of the German historical school, while Jules Michelet's rediscovery of Vico in 1824 started a new era in French writing on the Middle Ages.

HISTORIOGRAPHY IN THE 19TH AND 20TH CENTURIES

From the early 19th century, historiography began to develop in a radically different way. The decisive changes occurred among the German historians, largely through a reaction to the French Revolution and to a temporary subjugation of their country by Napoleon. Organized teaching of history in schools and universities became a matter of national importance, first in Prussia and then in other parts of Germany. As universal education spread to most European countries in the course of the 19th century, history was accepted everywhere as a necessary subject in schools. For the first time the bulk of historical writing came to be done by professional historians, for whom it became a condition of securing academic appointments or of consolidating their standings as university teachers. Historiography eventually became a continuously cooperative venture, where the achievements of past historians could be used systematically by their successors. But the growth of specialization and the bewildering number of types of works that came to be published constituted a new danger. In the past, important discoveries were frequently lost through lack of interest. But, by the second half of the 20th century, discoveries were in danger of being simply overlooked amid the flood of publications.

Growth of specialization

Another great change lay in the growth of intellectual freedom. Free expression of independent or unorthodox ideas had become dangerous during the French Revolution and under Napoleon, both in the territories controlled by the French and, by way of frightened reaction, in the lands of their unconquered opponents. After 1815 conditions for freer historiography improved gradually in much of Europe. Charles Darwin's *Origin of Species* (1859), which put forth a theory of evolution at first unacceptable to church authorities, probably could not have been published with the same impunity any earlier.

One feature of the growing tolerance of governments toward historiography was the gradual creation of public archives, such as the British Public Record Office in London, created in 1838, and the freer opening of the collections already in existence. Even the papacy accepted these changes, and Pope Leo XIII opened up the papal archive in 1883 as part of a deliberate new policy of encouraging historical study of Catholicism. For the first time historiography came to be based largely on unpublished records, and scholars were tempted into excessive reliance on original documents while unduly neglecting the older types of narrative sources.

In the 20th century some grievous threats to the persistence of free scholarship recurred, and historiography suffered with other branches of humane studies. The establishment of a Communist regime in Russia led, at first, to the rejection of most pre-1917 history as a fit subject for schools and universities. This decision was reversed in the 1930s, and from 1945 Communist countries were encouraging a form of historiography especially concerned with economic history and the class struggles of the past. There was also an enthusiastic interest in the material remains of past ages, leading to an impressive development of archaeology, particularly in Poland. The rise of dictatorships in Italy and Germany had disastrous effects on historiography in those countries, and recovery after World War II was only gradual.

Judged merely by the number of "practicing" historians and of their publications, historiography seemed in a very flourishing state in the 1970s. Its European traditions had spread to all the other continents and were largely accepted in all non-Communist countries.

The *Introduction aux études historiques* (*Introduction to the Study of History*) of Charles V. Langlois and Charles Seignobos (1898), supplemented by critical comments of another outstanding French historian, Ferdinand Lot (in *Le Moyen Age*, 1898), provides an excellent starting point for the discussion of modern historical methods. History is an autonomous branch of learning, and some of its methods may be unique. Historians should not try to formulate general laws; their branch of learning merely "aims at explaining reality." Langlois and Seignobos particularly stress that history is not a science of observation but a science of reasoning how to extract from imperfect documentary or narrative records some glimpses of what actually happened.

The historian's task

A historian has to subject his sources to a whole series of preliminary investigations. First comes "external criticism," aimed at determining whether the sources are appropriate and adequate for the particular task in hand. The provenance, date, and authenticity of each source must be established by using the techniques of diplomatic, the detailed study and assessment of documents, and of paleography, the study of ancient handwriting, and of other auxiliary sciences that were elaborated after the 17th century. In France a special institution for teaching some of these techniques, the École des Chartes, was created in 1821. The first specialized seminar for instruction in these subjects was established in 1854 at Vienna by Theodor von Sickel, one of the greatest medievalists of the 19th century, and it was gradually imitated by leading German universities. One of the most important critical refinements introduced in the course of the 19th century was the improved handling of narrative sources brought about by seeking to discover the literary sources that lay behind them. Leopold von Ranke, one of the foremost German historians, who began his career as a teacher of classics, was gradually attracted to history through a desire to understand better the sources of the Greek and Latin authors whom he was expounding. In the later decades of the 19th century, such a quest became a normal feature of historical scholarship.

Once a historian has decided, through the application of "external criticism," on the sources that are relevant to his purpose, he must next, by "internal criticism," make sure that he fully understands what he has selected. German classical philologists were the first to bring these latter investigations to a high degree of perfection. Karl Lachmann, an editor of the Latin poets, is justly regarded as the creator of modern textual criticism in its most rigorous forms, and historians gradually adopted similar methods. The language of the sources must be understood, corruptions in the text must be eliminated, and the historian must, as accurately as possible, penetrate the minds of the authors with whom he is dealing.

All these critical operations on the sources are merely preliminaries, and the work of the historian proper only starts when he attempts a synthesis of his materials. F. Lot stresses that in this qualities other than the erudite skills come into play. There must be sympathy with the subjects under study, for without it there can be no imaginative insight into the past. Ideally, a historian must display capacities akin to those of a poet or an artist.

Such a quality was, by and large, lacking in the work of the historians of the Enlightenment, who had been unable to achieve imaginative insight into civilizations very different from their own. The greatest shortcoming of Gibbon was his temperamental inability to appreciate religion. The new historiography of the 19th century was created chiefly by Germans, who, through a reaction to the ungodly and cosmopolitan Enlightenment, were endowed to excess with a passion for extolling the unique nature of their fatherland and for tracing the roots of this

uniqueness through the whole course of German history. These developments in German historiography can be traced back to some strands of German thought in the 18th century, especially to some features of the writings of Johann Gottfried von Herder. He denied that the purpose of history was to provide a bird's-eye view of the progress of the human mind. It was, rather, to reconstruct history as it had been, which means that all countries and periods are equally deserving of study. This view anticipated Ranke's oft-quoted aim to describe what has actually happened and his conviction that the description of all human history displays the workings of God's providence. The disasters inflicted upon Germany by Napoleon brought forth a patriotic school of historians whose urgent task it became to propagate these views as a means of restoring German independence. The centre of this movement was in Prussia, at the newly founded University of Berlin (1809). Wilhelm von Humboldt, its effective founder, believed that the task of the historian lay in discovering the ideas behind the facts. The concepts that had special validity for him were ideas of religion and of a national state. The German historical school prided itself on the scientific precision of its methods, on its determination to get all the details right, and on the scrupulous quotation of sources. This display of exact scholarship represented a great gain for historical sciences, but its chief purpose was to convince the reader. Yet these German historians were fundamentally inspired by a prejudiced, arbitrary set of assumptions. It is particularly difficult to detect Ranke's hidden bias, as he made a parade of refusing to pass judgments on the past. His preference for the study of foreign relations between states and his treatment of states as natural entities with a right to fulfill their individual destinies justified the successes of Prussia. The defeat in 1848 of the German aspirations to national unity inspired his pupil Wilhelm von Giesebrecht to write the history of the medieval German empire to remind his countrymen of their past glories. When German unification was achieved in 1871, Giesebrecht doubted whether there was any need to bring out any further volumes of his great work. But many German historians, having contributed mightily to the unification of Germany, continued to describe complacently the triumphs of the Bismarckian state. This was one of the purposes of the school of historical economists led by Gustav von Schmoller. There were some dissenting voices. Theodor Mommsen, the greatest historian of antiquity produced by the 19th century, deplored the tendency of his countrymen to worship state power. Friedrich Meinecke, a leading German historian of political ideas, who until 1914 accepted the ordinary nationalistic assumptions of his countrymen, gradually entirely changed his views and, after the defeat of Germany in two world wars, pleaded in his *Deutsche Katastrophe* (1946; *The German Catastrophe*) for a historiography concerned with the higher values of general civilization. Among the German historians, particularly striking progress was achieved in medieval studies. Meanwhile, attempts at imaginative reconstructions of the past were being made in other countries of western Europe. Jules Michelet wrote in 1833–43 the first history of medieval France based on the French national archives, of which he was at that time keeper. Macaulay's *History of England* (1848–61), covering chiefly the years 1685–1702, represented again a remarkable though prejudiced attempt to relive the past.

German patriotic historians

German scholarly techniques and the methods of German historical teaching spread to other countries in the course of the later 19th century, though it is important to note that until 1914 a significant proportion of leading historians from states outside Germany spent some time in that country. This is particularly true of some of the greatest Russian scholars, such as M.I. Rostovtzeff, one of the most important modern historians of antiquity. In England, William Stubbs, though self-taught, applied the results of German scholarship to the reconstruction of English medieval history. Gabriel Monod, who had studied in Germany, was prominent in introducing more scientific techniques into medieval French historiography, and he founded in 1876 the *Revue Historique* as the main organ of French historical scholarship. A succession of American students went to Germany, and some, on their return home, reorganized historical studies. Measured by the sheer bulk of publications, the amount of American history written since the 18th century is probably greater than that of any other modern nation. But apart from editions of sources, very few works on American history published before about 1900 are of much practical use today. The most influential pioneer in organizing scientific historiography was Herbert Baxter Adams, who between 1876 and his death in 1901 made the Johns Hopkins University at Baltimore into the foremost American centre of historical studies. He was also one of the founders of the American Historical Association in 1884 and played a large part in successfully launching the *American Historical Review* in 1895 as the main organ of historical scholarship. Some of Adams' pupils became great scholars in various fields of general history. Charles Homer Haskins' works on Norman institutions and on science and culture in the 12th and 13th centuries made him one of the foremost medievalists of the 20th century. But a movement for creating a purely American history was launched in 1893 by another of Adams' pupils, Frederick Jackson Turner, who inaugurated a "progressive" school of historians through his conviction that the fundamental fact of American history down to 1890 was the settlement of a continent. In Turner's eyes the main theme of American history in the 19th century was the conflict between the patrician and capitalist groups of the Eastern Seaboard and the needs of the new settlers in the Middle West. Charles A. Beard inaugurated by his *Economic Interpretation of the American Constitution* (1913) an attempt to rewrite the entire history of the U.S. in terms of conflicts between different groups of economic interests. The weakness of this type of historiography was that it encouraged an excessive parochialism. After 1945 the "progressive" historians came under fire both from more conservative scholars who preferred to stress elements of common tradition and purpose in American development and from the historians of the "new left." In the 1960s and 1970s the close connection between writings on American history and the active political life was infusing great variety and vitality into its historiography, though making it perhaps too susceptible to rapidly changing external pressures.

The work of Henry Baxter Adams

(E.B.Fr./Ed.)

Methodology of historiography

The methodology of history does not differ in broadest outline from that of other disciplines in its regard for existing knowledge, its search for new and relevant data, and its creation of hypotheses. It is the same for all historical writing, success depending on skill and experience: and division of the past on temporal or topical lines merely reflects the human limitations of historians. Although historical methodology has four facets, the more skilled the historian the less he gives them conscious consideration; and any historian is likely to be concerned with two or more concurrently. The four facets are heuristic, knowledge of current interpretation, research, and writing.

The first two may be briefly considered. Heuristic has been adopted as a convenient term for the technique of investigation that can be acquired solely by practice and experience. In the case of the historian it embraces such things as knowledge of manuscript collections, methods of card indexing and classifying material, and knowledge of bibliography. It underlies other aspects of methodology as in knowledge of the capabilities of historians working in the same and similar fields or in the power of dealing expeditiously with documentary material. The necessity for knowledge of current interpretation is based on the working principle that inquiry proceeds from the known to the unknown; and the historian has to be well acquainted with existing work in his own field, in contiguous historical fields and in allied disciplines. The work in each case consists of both "fact" and interpretation, and the amount the historian accepts will vary. In his own field he will normally not accept facts, and certainly not current interpretation, on trust; in contiguous historical fields he will accept facts and current interpretation by experts in those

fields, but qualified by heuristic and his general historical knowledge; in allied disciplines, such as anthropology, economics, geography, natural science, philology, psychology, sociology, he must unless there is strong evidence to the contrary presume the technical skill and intellectual honesty of scholars in those fields. There is, of course, no reason why a historian cannot be reasonably versed in one or more of these and other disciplines, and should the nature of his enquiry demand it he must be.

Historical research is the term applied to the work necessary for the establishing of occurrences, happenings, or events in the field with which the historian is concerned. Knowledge of these is entirely dependent on the transmission of information from those living at the time, and this information forms what is known as the source material for the particular period or topic. The occurrences themselves can never be experienced by the historian, and what he has at his disposal are either accounts of occurrences as seen by contemporaries or something, be it verbal, written, or material, that is the end product of an occurrence. These accounts or end products have been variously termed relics, tracks, or traces of the occurrences that gave rise to them; and from them the historian can, with varying degrees of certainty, deduce the occurrences. The traces are thus the "facts" of history, the actual occurrences deductions from the facts; and historical research is concerned with the discovery of relevant traces and with deduction from those traces insofar as this will aid the search for further relevant traces.

SOURCE MATERIAL

Source material falls into three groups which can be differentiated as written, material, and traditional. Written source material has two subdivisions, literary, sometimes called subjective, and official. The first consists of events as seen through the eyes of an individual and therefore as interpreted by him, normally entailing selection of occurrences or attribution of motive. The second subdivision, the official, consists of records produced in transacting business at any level from individual to international. The information given is basically in statement form, impersonal, and containing only the most superficial suggestions of causation and motivation. In practice the boundary between literary and official sources is blurred and a document may contain elements of both. The second main division, material source material, consists of objects that have resulted from activities of human beings in the past. The third group, traditional source material, covers what is handed on verbally or as practices, although later generations may commit such things to writing. Obvious examples are archaic forms, traditional practices, nursery rhymes, folklore, and place names. Comparison with parallel source material and knowledge of current interpretation will normally show the historian whether his particular source can be presumed true, partially true, or faked. If true or partially true allowance has to be made for the subjective element in literary and some traditional sources and for the difficulty of reconstructing the events themselves from the traces surviving in official, material, or other traditional sources.

The classification of source material is essentially pragmatic, based on the differing techniques required in handling sources of the different groups: an inscribed tombstone, for example, can be either a written or a material source depending on whether the historian's concern is with the content of the inscription or with the stone. Specialized training in what are sometimes known as ancillary disciplines may, depending on the nature of his investigation, be necessary for the historian. The most important of these are archaeology, bibliography, chronology, diplomatics, epigraphy, genealogy, paleography, sigillography, and textual criticism. It need hardly be said that the historian must have competence in the languages used in his source material. Many historians give part of their time to the editing of source material. This is not historical writing but is of use to other historians in the same field. The collection of facts as an end in itself is, however, antiquarianism not history, and the essential end product of historical investigation is the historian's own writing.

USING SOURCE MATERIAL

The question of what history is belongs to the philosophy rather than the methodology of history. The word history itself is used ambiguously to describe both the past and what is written about the past; but it is this second meaning that is relevant to the working definition that history is the past experience of society. For what reasons society may wish to utilize its past experience is not the concern of the historian, whose task is to make available to society that past experience and to record it for future reference. An individual utilizing his own past experience has to recall the significant elements of that experience with accuracy and establish their causal and chronological relationships. The historian behaves similarly concerning the past experiences of society; but the reconstruction of events from traces, the selection of those relevant to his task, and the establishing of relationships allow a varying freedom of choice by the historian, which thus introduces the subjective element of the historian's personality. This cannot be eliminated from historical writing, and the historian's aim is to make the margin of intellectual error as small as possible. The handling of source material demands only care and technical competence, and it is mainly in the construction of hypotheses and in the establishing of relationships that this intellectual error can enter. A check is provided by the opinions of other historians working in the same field. His work will, if accepted, become part of current interpretation, sometimes described as accepted history but, as with all current interpretation, subject to revision by himself or others.

Historical methodology became more clearly formulated during the 19th and 20th centuries, but there have been historians at times long past whose work can be judged by present-day standards. There are, however, certain important differences between present methodology and the general run of past methodology. Much medieval writing, for example, bows to precedent in literary sources and in current interpretation, and uncritical acceptance of an earlier writer's work can occur century after century. The comparative neglect of official sources by the majority of European historians before the 19th century gave no corrective to literary sources. The greatest impediment to the development of modern methodology lay, however, in the varying concepts of history, some of which survive today. The concept of history as a form of literature made it a type of imaginative art on which judgment was passed on grounds of elegance rather than accuracy. Closely allied with this is the "ethical" concept of history whereby historical writing became a series of value judgments on individuals and actions. The converse of this was the impossible "objective" or "scientific" history of the later 19th century, though it did popularize the concept of research and developed the ancillary disciplines. The use of history for propaganda purposes is in its crudest form virtually a branch of fiction and thus independent of research; in its more subtle forms it can encourage accuracy in research, but it will encourage also the suppression of inconvenient traces and intellectual dishonesty in the elucidation of relationships. In this it indicates one of the main impediments to methodological and historical development: the holding by the historian of *a priori* theories or laws to which all events and relationships must conform, whether it be the theory of divine intervention in human affairs favoured in medieval times or the Marxist theory current over much of the modern world.

(Ed.)

Ancillary fields

ARCHAEOLOGY

The word archaeology comes from the Greek *archaia* ("ancient things"), and *logos* ("theory" or "science"). It has been used in varying ways, but from the late 18th century onward has come to mean that branch of learning that studies the material remains of man's past. This includes man's artifacts, from the very earliest stone tools, of perhaps 2,000,000 years ago, to the man-made objects that are buried or thrown away at the present day: everything made by human beings—from simple tools to complex

machines, from the earliest houses and temples and tombs to palaces, cathedrals, and pyramids.

The archaeologist is first a descriptive worker: he has to describe, classify, and analyze the artifacts he studies. An adequate and objective taxonomy is the basis of all archaeology, and many good archaeologists spend their lives in this activity of description and classification. But the main aim of the archaeologist is to place the material remains in historical contexts, to supplement what may be known from written sources, and, thus, to increase understanding of the past. Ultimately, then, the archaeologist is a historian: his aim is the interpretive description of the past of man.

Increasingly, many scientific techniques are used by the archaeologist, and he uses the scientific expertise of many persons who are not archaeologists in his work. The artifacts he studies must often be studied in their environmental contexts; and botanists, zoologists, soil scientists, and geologists may be brought in to identify and describe plants, animals, soils, and rocks. Radioactive carbon dating, which has revolutionized much of archaeological chronology, is a by-product of research in atomic physics. But although archaeology uses extensively the methods, techniques, and results of the physical and biological sciences, it is not a natural science; some consider it a discipline that is half science and half humanity. Perhaps it is more accurate to say that the archaeologist is first a craftsman, practicing many specialized crafts (of which excavation is the most familiar to the general public), and then a historian.

The nature of archaeological work

The justification for this work is the justification of all historical scholarship: to enrich the present by knowledge of the experiences and achievements of our predecessors. Because it concerns things people have made, the most direct findings of archaeology bear on the history of art and technology; but by inference it also yields information about the society, religion, and economy of the people who created the artifacts. Also, it may bring to light and interpret previously unknown written documents, providing even more certain evidence about the past.

But no one archaeologist can cover the whole range of man's history, and there are many branches of archaeology divided by geographical areas (such as classical archaeology, the archaeology of ancient Greece and Rome, or Egyptology, the archaeology of ancient Egypt) or by periods (such as medieval archaeology and industrial archaeology). Writing began 5,000 years ago in Mesopotamia and Egypt; its beginnings were somewhat later in India and China, and later still in Europe. The aspect of archaeology that deals with the past of man before he learned to write has, since the middle of the 19th century, been referred to as prehistoric archaeology, or prehistory. In prehistory the archaeologist is paramount, for here the only sources are material and environmental.

The scope of this article is to describe briefly how archaeology came into existence as a learned discipline; how the archaeologist works in the field, museum, laboratory, and study; and how he assesses and interprets his evidence and transmutes it into history.

History of archaeology. No doubt there have always been people who were interested in the material remains of the past, but archaeology as a discipline has its earliest origins in 15th- and 16th-century Europe, when the Renaissance Humanists looked back upon the glories of Greece and Rome. Popes, cardinals, and noblemen in Italy in the 16th century began to collect antiquities and to sponsor excavations to find more works of ancient art. These collectors were imitated by others in northern Europe who were similarly interested in antique culture. All this activity, however, was still not archaeology in the strict sense. It was more like what would be called art collecting today.

The Mediterranean and the Middle East. Archaeology proper began with an interest in the Greeks and Romans and first developed in 18th-century Italy with the excavations of the Roman cities of Pompeii and Herculaneum. Classical archaeology was established on a more scientific basis by the work of Heinrich Schliemann, who investigated the origins of Greek civilization at Troy and Mycenae in the 1870s; of M.A. Biliotti at Rhodes in this same period; of the German Archaeological Institute under Ernst Curtius at Olympia from 1875 to 1881; and of Alexander Conze at Samothrace in 1873 and 1875. Conze was the first person to include photographs in the publication of his report. Schliemann had intended to dig in Crete but did not do so, and it was left to Arthur Evans to begin work at Knossos in 1900 and to discover the Minoan civilization, ancestor of classical Greece.

Egyptian archaeology began with Napoleon's invasion of Egypt in 1798. He brought with him scholars who set to work recording the archaeological remains of the country. The results of their work were published in the *Description de l'Égypte* (1808–25). As a result of discoveries made by this expedition, Jean-François Champollion was able to decipher ancient Egyptian writing for the first time in 1822. This decipherment, which enabled scholars to read the numerous writings left by the Egyptians, was the first great step forward in Egyptian archaeology. The demand for Egyptian antiquities led to organized tomb robbing by men such as Giovanni Battista Belzoni. A new era in systematic and controlled archaeological research began with the Frenchman Auguste Mariette, who also founded the Egyptian Museum at Cairo. The British archaeologist Flinders Petrie, who began work in Egypt in 1880, made great discoveries there and in Palestine during his long lifetime. Petrie developed a systematic method of excavation, the principles of which he summarized in *Methods and Aims in Archaeology* (1904). It was left to Howard Carter and Lord Carnarvon to make the most spectacular discovery in Egyptian archaeology, that of the tomb of Tutankhamen in 1922.

The development of Egyptology

Mesopotamian archaeology also began with hectic digging into mounds in the hopes of finding treasure and works of art, but gradually these gave way in the 1840s to planned digs such as those of the Frenchman Paul-Émile Botta at Nineveh and Khorsabad, and the Englishman Austen Henry Layard at Nimrud, Kuyunjik, Nabi Yunus, and other sites. Layard's popular account of his excavations, *Nineveh and Its Remains* (1849), became the earliest and one of the most successful archaeological best-sellers. In 1846 Henry Creswicke Rawlinson became the first man to decipher the Mesopotamian cuneiform writing. Toward the end of the 19th century, systematic excavation revealed a previously unknown people, the Sumerians, who had lived in Mesopotamia before the Babylonians and Assyrians. The most impressive Sumerian excavation was that of the Royal Tombs at Ur by Leonard Woolley in 1926.

First steps to archaeology. The development of scientific archaeology in 19th-century Europe from the antiquarianism and treasure collecting of the previous three centuries was due to three things: a geological revolution, an antiquarian revolution, and the propagation of the doctrine of evolution. Geology was revolutionized in the early 19th century with the discovery and demonstration of the principles of uniformitarian stratigraphy (which determines the age of fossil remains by the stratum they occupy below the earth) by men like William Smith, Georges Cuvier, and Charles Lyell. Lyell, in his *Principles of Geology* (1830–33), popularized this new system and paved the way for the acceptance of the great antiquity of man. Charles Darwin regarded Lyell's *Principles* as one of the two germinal works in the formation of his own ideas on evolution. Early stone tools had been identified in Europe since mid-16th century. That they were, however, older than 4004 BC, the date of man's origin according to biblical chronology, was not recognized until late in the 18th century, when John Frere suggested a great age for artifacts found in Suffolk, England, based on their location in certain strata. The discoveries of Jacques Boucher de Perthes in the Somme Valley in France, and of William Pengelly in the caves of South Devon in England, were used to demonstrate the antiquity of man in 1859, the same year that saw the publication of Darwin's revolutionary *Origin of Species*. Approximate dates for the Paleolithic Period (Old Stone Age) of the prehistoric past were thus established, although the expression "Palaeolithic" was not used until John Lubbock coined it in his book *Pre-historic Times* (1865).

Half a century before this, Scandinavian archaeologists had created a revolution in antiquarian thought by postulating, on archaeological grounds, successive technological stages in man's past. C.J. Thomsen classified the material in the Copenhagen Museum, opened to the public in 1819, on the basis of three successive ages of Stone, Bronze, and Iron. His pupil and successor, J.J.A. Worsaae, showed the correctness of this museum arrangement by observed stratigraphy in the Danish peat bogs and barrows (funerary mounds). Low lake levels in Switzerland in the mid-1850s permitted the excavation of the prehistoric Swiss lake dwellings, and here again the theory of a succession of technological stages was confirmed.

Darwin's *Origin of Species* implied a long past for man, and the acceptance of the idea of human evolution in the last four decades of the 19th century created a climate of thought in which archaeology flourished and that led to great advances in the unfolding of the full story of man's development.

In his *Pre-historic Times*, Lubbock expanded the three-age system of Thomsen and Worsaae to a four-age system, dividing the Stone Age into Old and New periods (Paleolithic and Neolithic). In the last quarter of the 19th century remarkable Paleolithic discoveries were made in France and Spain; these included the discovery and authentication of actual works of sculpture and cave paintings from the Upper (later) Paleolithic Period (*c.* 30,000–*c.* 10,000 BC). When Marcellino de Sautuola discovered the cave paintings at Altamira, Spain (1875–80), most experts refused to believe they were Paleolithic; but after similar discoveries at Les Eyzies in France around 1900, they were accepted as such and were recognized as one of the most surprising and exciting archaeological discoveries. A succession of similar finds has continued in the 20th century. The most famous of these was at Lascaux, France, in 1940.

During the last quarter of the 19th century, Gen. A.H. Pitt-Rivers' excavations of prehistoric and Roman sites at Cranborne Chase, Dorset, laid the foundations of modern scientific archaeological field technique, which was later developed and improved in England and Wales by men such as Sir Mortimer Wheeler and Sir Cyril Fox.

20th-century developments. The 20th century has seen the extension of archaeology outside the areas of the Near East, the Mediterranean, and Europe, to other parts of the world. In the early '20s, excavations at Mohenjo-Daro and Harappā, in present Pakistan, revealed the existence of the prehistoric Indus civilization. In the late '20s, excavations at An-yang in eastern China established the existence of a prehistoric Chinese culture that could be identified with the Shang dynasty of early Chinese records.

Studies of the Stone Age

The Stone Age has been described and studied throughout the world; among the most sensational discoveries are those of L.S.B. Leakey, who found stone tools and skeletal remains of early man dating back 2,000,000 years in the Olduvai Gorge in Tanzania. Intensive work of great importance has brought to light early Neolithic sites at Jericho in Palestine; Hassuna, Iraq; Çatalhüyük, Turkey; and elsewhere in the Near East, establishing the origins of agriculture in that region.

Serious archaeological work began later in America than Europe, but as early as 1784 Thomas Jefferson had excavated mounds in Virginia and made careful stratigraphical observations. The 20th century has seen a great increase in archaeological knowledge about prehistoric America: two startling advances were the discovery of the origin of domesticated crops (including maize) in Central America and of the Olmec civilization of Mexico (1000–300 BC)—the oldest of the New World civilizations and probably the parent of all the others.

The enormous growth of archaeological work has meant the establishment of archaeology as an academic discipline; few important universities anywhere in the world are now without professors and departments of archaeology. There are now a very large number of scholarly journals in the field, as well as a considerable body of popularized books and journals that attempt to bridge the gap between professional and layman.

Fieldwork. *Preliminary work.* Some archaeologists call everything they do out-of-doors fieldwork, but others distinguish between fieldwork, in a narrower sense, and excavation. Fieldwork, in the narrow sense, consists of the discovery and recording of archaeological sites and their examination by methods other than the use of the spade and the trowel. Sites hitherto unknown are discovered by walking or motoring over the countryside: deliberate reconnaissance is an essential part of archaeological fieldwork.

In Europe, a study of old records and place-names may lead to the discovery of long-forgotten sites. The mapping of new and old sites is an essential part of archaeological survey. This process has been brought to a very high standard of perfection, both in the marking of archaeological sites on ordinary topographical maps and in the production of special period maps. The distribution map of artifacts, especially when studied against the background of the natural environment, is a key method of archaeological study.

Aerial photography in archaeology

The formerly earthbound archaeologist has been greatly helped by the development of aerial photography. The application of aerial photography to archaeological investigation began in a small way during World War I, as a side effect of military reconnaissance, and was given further impetus by World War II; the photographic intelligence departments of all the combatant nations were extensively staffed by archaeologists, who then carried their expertise and enthusiasm into the postwar years. The University of Cambridge now has its own department of air photography under J.K.S. St Joseph: using its own pilot and aircraft, it flies photographic missions over Ireland, Great Britain, Denmark, and The Netherlands. The number of new sites discovered each year by aerial photography is very large. Some of these are surface sites, especially partly destroyed sites that show up well in special conditions of light, as in early morning or late evening. But many are sites that could not be found on the ground and that show up in aerial photographs as variations in soil colour or in the density of crop.

Archaeological reconnaissance may be advanced from ordinary surface or aerial methods in a wide variety of ways. A very simple method is tapping the ground to sound for substructures and inequalities in the subsoil. Deep probes have made it possible to trace walls and ditches. The Lerici Foundation of Milan and Rome has had great success with this method since its development of the Nistri periscope, first used in 1957 in an Etruscan tomb in the cemetery of Monte Abbatone. The periscope is inserted into the burial chamber and can photograph the walls and contents of the whole tomb.

Other modern techniques that have been applied to archaeological prospecting employ electricity and magnetic fields (geophysical prospecting). A method of electrical prospecting had been developed in large-scale oil prospecting: this technique, based on the degree of electrical conductivity present in the soil, began to be used by archaeologists in the late 1940s and has since proved very useful. Magnetic methods of prospecting detect buried features by locating the magnetic disturbances they cause: these were introduced in 1957–58 and use such machines as the proton magnetometer, the proton gradiometer, and the fluxgate gradiometer. An American expedition discovered the site of Sybaris in Sicily by magnetic prospecting. Electromagnetic methods have been in use only since 1962; they employ developments of the concepts used in mine detectors. Instruments such as the pulsed-induction meter and the soil-conductivity meter detect magnetic soil anomalies, but only if the features are fairly shallow.

Excavation. Excavation is the surgical aspect of archaeology: it is surgery of the buried landscape and is carried out with all the skilled craftsmanship that has been built up in the last hundred years since Schliemann and Flinders Petrie. Excavations can be classified, from the point of view of their purpose, as planned, rescue, or accidental. Most important excavations are the result of a prepared plan—that is to say, their purpose is to locate buried evidence about an archaeological site. Many are project oriented: as, for example, when a scholar studying the life of the pre-Roman, Celtic-speaking Gauls of France may deliberately select a group of hill forts and excavate them,

Archaeology in urban areas

as Sir Mortimer Wheeler did in northwestern France in the years before the outbreak of World War II. But many excavations, particularly in the heavily populated areas of central and northern Europe, are done not from choice but from necessity. Gravel digging, clearing the ground for airports, quarrying, road widening and building, the construction of houses, factories, and public buildings frequently threaten the destruction of sites known to contain archaeological remains. Emergency excavations then have to be mounted to rescue whatever knowledge of the past can be obtained before these remains are obliterated forever. Partial destruction of cities in western Europe by bombing during World War II allowed rescue excavations to take place before rebuilding. A temple of Mithras in the City of London, Viking settlements in Dublin and at Århus, Denmark, and the original 6th-century-BC Greek settlement of Massalia (Marseille) were discovered in this way. An extension of the runways at London Airport led to the discovery of a pre-Roman Celtic temple there.

The role of chance in the discovery of archaeological sites and portable finds is considerable. Farmers have often unearthed archaeological finds while plowing their fields. The famous painted and engraved Upper Paleolithic cave of Lascaux in southern France was discovered by chance in 1940 when four French schoolboys decided to investigate a hole left by an uprooted tree. They widened a smaller shaft at the base of the hole and jumped through to find themselves in the middle of this remarkable pagan sanctuary. Similarly, the first cache of the Dead Sea Scrolls was discovered in 1947 by a Bedouin looking for a stray animal. These accidental finds often lead to important excavations. At Barnénès, in north Brittany, a contractor building a road got his stone from a neighbouring prehistoric cairn (burial mound) and, in so doing, discovered and partially destroyed a number of prehistoric burial chambers. The French archaeologist P.-R. Giot was able to halt these depredations and carry out scientific excavations that revealed Barnénès to be one of the most remarkable and interesting prehistoric burial mounds in western Europe.

All forms of archaeological excavation require great skill and careful preparation. Years of training in the field, first as an ordinary digger, then as a site supervisor, with spells of work as recorder, surveyor, and photographer, are required before anyone can organize and direct an excavation himself. Most museums, universities, and government archaeological departments organize training excavations. The very words dig and digging may give the impression to many that excavation is merely a matter of shifting away the soil and subsoil with a spade or shovel; the titles of such admirable and widely read books as Leonard Woolley's *Spadework* (1953) and *Digging Up the Past* (1930) and Geoffrey Bibby's *Testimony of the Spade* (1956) might appear to give credence to that view.

Excavation method

Actually, much of the work of excavation is careful work with trowel, penknife, and brush. It is often the recovery of features that are almost indistinguishable from nonarchaeological aspects of the buried landscape: one example of this is the recovery of mud-brick walls in Mesopotamia; another is the tracing of collapsed walls of dry stone slabs in a cairn in stony country in the southwest Midlands of England. Sometimes it is the recovery of features of which only ghost traces remain, like the burnt-out bodies from the buried city of Pompeii, or the strings of a harp that were found among the furnishings of Mesopotamian tombs at Ur.

Because of the damage he may cause by inexperience and haste, the untrained amateur archaeologist often hinders the work of the professional. Amateur archaeology is forbidden in many countries by stringent antiquity laws. At the same time, it is certainly true that nonprofessionals have made important contributions in many areas of archaeology. Occasionally, an amateur does make an important discovery the further excavation of which can then be taken over by trained professionals. Such was the case at Sutton Hoo in Suffolk in 1939, when work begun by a competent amateur was taken over by a team of experts who were able to uncover a great Anglo-Saxon burial boat and its treasure, without doubt the most remarkable archaeological find ever made in Britain.

There are, of course, many different types of archaeological sites, and there is no one set of precepts and rules that will apply to excavation as a whole. Some sites, such as temples, forts, roads, villages, ancient cities, palaces, and industrial remains, are easily visible on the surface of the ground. Among the most obvious archaeological sites that have yielded spectacular results by excavation are the huge man-made mounds (tells) in the Near East, called in Arabic *tilāl*, and in Turkish *tepes* or *hüyüks*. They result from the accumulation of remains caused by centuries of human habitation on one spot. The sites of the ancient cities of Troy and Ur are examples. Another type consists of closed sites such as pyramids, chambered tombs, barrows (burial mounds), sealed caves, and rock shelters. In other cases there are no surface traces, and the outline of suspected structures is revealed only by aerial or geophysical reconnaissance as described above. Finally, there are sites in cliffs and gravel beds, where many Paleolithic finds have been made.

The wide range of techniques employed by the archaeologist vary in their application to different kinds of sites. The opening of the tomb chamber in an Egyptian pyramid is, for example, a very different operation from the excavation of a tell in Mesopotamia or a barrow grave in western Europe. Some sites are explored provisionally by sampling cuts known as sondages. Large sites are not usually dug out entirely, although a moderate-sized round barrow may be completely moved by excavation. Whatever the site and the extent of the excavation, one element of the technique is common to all digs, namely, the use of the greatest care in the actual surgery and in the recording of what is found by word, diagram, survey, and photography. To a certain extent all excavation is destruction, and the total excavation of a site subsequently engulfed by a housing estate or gravel digging is total destruction. This is why the archaeologist's field notes and his published report become primary archaeological documents. They are not themselves, strictly speaking, archaeological facts: they are the excavator's interpretation of what he saw, or thought he saw, but this is the nearest the discipline can ever get to archaeological facts as established by excavation. The really great excavators leave such a fine record of their digs that subsequent archaeologists can re-create and reinterpret what they saw and found. To delay publishing the results of an excavation within a reasonable time is a serious fault from the point of view of archaeological method. An excavation is not complete until the printed report is available to the world. Often the publication of the report takes as long as, or much longer than, the actual work in the field.

Publication of results

When a site like the Palace of Minos at Knossos or the city of Harappā in Pakistan has been excavated, and the excavations are over, the excavator and the antiquities service of the country concerned have to face the problem of what to do with the excavated structures. Should they be covered in again, or should they be preserved for posterity, and if preserved, what degree of conservation and restoration is permissible? This is the same kind of problem that arises in connection with the removal of antiquities from their homeland to foreign museums, and there is no generally accepted answer to it. These problems remain to beset archaeology: should Sir Arthur Evans have reconstructed the Palace of Minos at Knossos? Should the art treasures of ancient Greece and Egypt, now in western European museums, be returned? There is no simple, straightforward, overall answer to these difficult questions.

Underwater archaeology. Underwater archaeology is a branch of reconnaissance and excavation that has been developed only during the 20th century. It involves the same techniques of observation, discovery, and recording that are the basis of archaeology on land, but adapted to the special conditions of working underwater. It is obvious that no archaeologist working on submarine sites can get far unless he is trained as a diver. Helmeted sponge divers have made most of the important archaeological discoveries in the Mediterranean. The French scientist Jacques-Yves Cousteau developed the self-contained breathing ap-

paratus known as the scuba, of which the most commonly used type is the aqualung. Cousteau's work at Le Grand Congloué near Marseille was a pioneer underwater excavation, as was the work of the Americans Peter Throckmorton and George Bass off the coast of southern Turkey. In 1958 Throckmorton found a graveyard of ancient ships at Yassı Ada and then discovered the oldest shipwreck ever recorded, at Cape Gelidonya—a Bronze Age shipwreck of the 14th century BC. George Bass of the University of Pennsylvania worked on a Byzantine wreck at Yassı Ada from 1961 onward, developing the mapping of wrecks photogrammetrically with stereophotographs and using a two-man submarine, the "Asherah," launched in 1964. The "Asherah" was the first submarine ever built for archaeological investigation.

Interpretation. Excavation often seems to the general public the main and certainly the most glamorous aspect of archaeology; but fieldwork and excavation represent only a part of the archaeologist's work. The other part is the interpretation in cultural and historical contexts of the facts established—by chance, by fieldwork, and by digging—about the material remains of man's past. This task of interpretation has five main aspects.

Classification and analysis. The first concern is the accurate and exact description of all the artifacts concerned. Classification and description are essential to all archaeological work, and, as in botany and zoology, the first requirement is a good and objective taxonomy. Second, there is a need for interpretive analysis of the material from which artifacts were made. This is something that the archaeologist himself is rarely equipped to do; he has to rely on colleagues specializing in geology, petrology (analysis of rocks), and metallurgy. In the early 1920s, H.H. Thomas of the Geological Survey of Great Britain was able to show that stones used in the construction of Stonehenge (a prehistoric construction on Salisbury Plain in southern England) had come from the Prescelly Mountains of north Pembrokeshire; and he established as a fact of prehistory that over 4,000 years ago these large stones had been transported 200 miles from west Wales to Salisbury Plain. Detailed petrological analysis of the material of Neolithic polished stone axes have enabled archaeologists to establish the location of prehistoric ax factories and trade routes. It is also now possible, entirely on a petrological basis, to study the prehistoric distribution of obsidian (a volcanic glass used to make primitive tools).

Petrological analysis

In the third place, the archaeologist, having dealt with the material of his artifacts by classification and taxonomy, and with its physical nature by petrology and metallurgy, turns to the remaining information he can get from his colleagues in the natural sciences. These tell him the environmental conditions in which the people he is studying lived; he now sees his material remains not as isolated artifacts but in the context of their original environments.

Dating. Having analyzed his discoveries according to their form, material, and biological association, the archaeologist then comes to the all-important problem of dating. Many material remains of man's past have no dating problem: they may be, like coins, or most coins, self-dating, or they may be dated by man-made dates in written records. But the great and difficult part of the archaeologist's work is dating material remains that are not themselves dated. This can be done in one of three ways. Sometimes an object from another culture, the date of which is known (*e.g.*, in the case of pottery, by its style), is found at a previously undated site. Then, using the relative dating principle (see below) the archaeologist reasons that the material found with the imported object is contemporary with it. Conversely, an object from an undated culture may be found at a site whose date is known. Thus nonliterate communities can be dated by their contact with literate ones. This technique is known as cross dating; it was first developed by Sir Flinders Petrie when he dated Palestinian and early Greek (Aegean) sites by reference to Egyptian ones. Much of the prehistoric chronology of Europe in the Neolithic, Bronze, and Early Iron ages is based on cross dating with the ancient Near East.

Aside from cross dating, the archaeologist faced with material in a site having no literate chronological evidence of its own has two other ways of dating his material. The first is relative, the second absolute. Relative dating merely means the relation of the date of anything found to the date of other things found in its immediate neighbourhood. As has already been described, this method also plays a part in cross dating. Stratigraphy is the essence of relative dating. The archaeologist observes the accumulation of deposits in a gravel pit, a peat bog, in the construction of a barrow, or in accumulated settlements in a tell, and, like the geologists who introduced the principles of stratigraphy in the late 18th and early 19th centuries, he can see the succession of layers in the site and can then establish the chronology of different levels of layers relative to each other. In the excavation of a great tell like Ur or Troy the relative chronology of the various levels of occupation is the first thing to be established. Some archaeologists, even until quite recent times, have mistakenly supposed that depth below ground level is itself an indication of antiquity.

Relative dating

But even in properly observed and recorded stratigraphic levels there is often doubt, and the question arises: are all the artifacts and human remains found in the same level contemporary? Is it possible that there could have been later intrusions that have been difficult to distinguish in the field? The analysis of the fluorine content of bones has been very helpful here. Recognized as a valuable technique by French scientists in the 19th century, it was developed in England by K.P. Oakley in the 1950s. If bones in apparently the same geological or archaeological level have markedly different fluorine content, then it is clear that there must be interference—for example, by a later burial, or by deliberate planting of faked remains, as happened in the case of the Piltdown "Man" hoax in England.

Absolute man-made chronology based on king lists and records in Egypt and Mesopotamia goes back only 5,000 years. For a long time archaeologists searched for an absolute chronology that went beyond this and could turn their relative chronologies into absolute dates. Clay-varve counting seemed to provide the first answer to this need for a nonhuman absolute chronology. Called geochronology by Baron Gerard De Geer, its Swedish inventor, this method was based on counting the thin layers of clay left behind by the melting glaciers when the European Ice Age came to an end. This gave a chronology of about 18,000 years—three times as long as the man-made chronology based on Egyptian and Mesopotamian king lists. Thus, absolute dates could be established for artifacts from the Late Paleolithic Period, the whole of the Mesolithic Period, or Middle Stone Age, and much of the Early Neolithic Period.

Dendrochronology, the dating of trees by counting their growth rings, was first developed for archaeological purposes by A.E. Douglass in the United States. The application of this method to archaeology depends, obviously, on the use in antiquity of old datable trees in the construction of houses and buildings. It has been possible by dendrochronology to date prehistoric American sites as far back as the 3rd and 4th centuries BC.

The greatest revolution in prehistoric archaeology occurred in 1948, when Willard F. Libby, at the University of Chicago, developed the process of radioactive carbon dating. In this method, the activity of radioactive carbon (carbon-14) present in bones, wood, or ash found in archaeological sites is measured. Because the rate at which this activity decreases in time is known, the approximate age of the material can be determined by comparing it to carbon-14 activity in presently living organic matter. There have been problems and uncertainties about the application of the radioactive carbon method, but, although it is less than perfect, it has given archaeology a new and absolute chronology that goes back 40,000 years.

Radioactive carbon dating

Following the revolutionary discovery of radioactive carbon dating, other physical techniques of absolute dating were developed, among them potassium–argon dating and dating by thermoluminescence. Potassium–argon dating has made it possible to establish that the earliest remains of man and his artifacts in East Africa go back at least 2,000,000 years, and probably further.

Historical judgments. The last and most important task

of the archaeologist is to transmute his interpretation of the material remains he studies into historical judgments. When he is dealing with medieval and modern history he is often doing no more than adding to knowledge already available from documentary sources: but even so his contribution is often of great importance; for example, in relation to the growth and development of towns and the study of deserted medieval villages. When he is dealing with ancient history and prehistory, he is making a contribution of the greatest importance and often one that is more important than that of purely literary and epigraphical sources. For the prehistoric period, which now appears to stretch from 2,000,000 years ago to about 3000 BC, archaeological evidence is the only source of knowledge about human activities. But prehistoric remains have always been the most difficult to interpret, precisely because there are no written records to aid in the task. Now, with exact dating techniques at his disposal, the prehistorian is becoming more like the historical archaeologist and is concerned with the periodization and the historical contexts of his finds. (G.E.D.)

BIBLIOGRAPHY

Bibliography, the art or science of the description of books, has acquired special importance in the 20th century because of the need for effective organization of the records of human communication in the face of the enormous growth of publishing activity and the need, especially in undeveloped countries, for informed access to the world's scientific and technical information. It has been said that without bibliography, the records of civilization would be an uncharted chaos of miscellaneous contributions to knowledge, unorganized and inapplicable to human needs.

The word bibliography, in its literal sense, derived from the Greek *bibliographia* (2nd century AD), means the writing of books, and it was so defined in the 17th century; since the 18th century, it has been used to denote the systematic description and history of books. It is now commonly used in two widely divergent, though basically connected senses: (1) the listing of books, arranged according to some system (in this sense it is called enumerative, systematic, or descriptive bibliography); and (2) the study of books as material objects; *i.e.*, the study of the material of which books are made and the manner in which they are put together (in this sense commonly called critical bibliography). It is the function of bibliography to provide useful information for the student, in the one case supplying him with information about material for study, in the other helping him to establish the place of a book (or a piece of writing) in an author's production and its quality and authenticity as a text for study.

Kinds of bibliographies

Descriptive bibliography. The tasks of the compiler of a bibliography are (1) to find out what books on a particular subject exist; (2) to describe them item by item; and (3) to assemble the resulting entries into useful arrangements for reference and study. The need for lists of this kind arises as soon as the number of books in any subject is too great to be easily remembered. Among the earliest lists of books are those compiled by certain writers as guides to their own books—*e.g.*, the celebrated physician Galen compiled, in the 2nd century AD, a description in subject order of his own writings (*De propriis libris liber*, "A Book About My Own Books"); the Venerable Bede attached at the end of his *Ecclesiastical History of the English People*, in 731, an autobiographical note with a list of his writings; Erasmus also published a catalog of his own writings in narrative form in 1523. Early bibliographies, however, were not confined to such autobibliographies: as early as the last decade of the 4th century AD, the idea of attaching lists of works to lives of ecclesiastical writers was adopted by St. Jerome in his *De viris illustribus* ("Concerning Famous Men").

The invention of printing in the 1440s made possible the rapid multiplication of books, thereby increasing the need for guides to the resultant literature and also bringing about changes in the manner of compilation. The earliest substantial bibliography after the invention of printing was that of Johannes Tritheim, abbot of Sponheim in the diocese of Mainz, who in *Liber de scriptoribus ecclesiasticis* ("Book About Ecclesiastical Writers," 1494) included in chronological order, with an alphabetical index, about 1,000 ecclesiastical writers and their books. In 1545 a German-Swiss writer and naturalist, Conrad Gesner, who has been known as the father of bibliography, published his *Bibliotheca Universalis* (*Universal Bibliography*) of all Latin, Greek, and Hebrew writers, living and dead; this was followed three years later by a second volume, *Pandectarum sive Partitionum universalium libri XXI* ("Twenty-one Books of Encyclopedias or Universal Divisions [of Knowledge]"), in which the entries, arranged alphabetically in the earlier volume, are rearranged under 21 subject headings. Although Gesner was not the earliest descriptive bibliographer, his attempts at universality and classification earn him his fame. Gesner's idea of universality remained an ideal until recent times; in the face of the formidable problems of size, cost, and complexity, however, it has receded into the background of human aspirations. The Institut International de Bibliographie, founded in 1895 in Brussels by Henry Lafontaine and Paul Otlet with the object of creating a universal bibliography of books and articles in periodicals, arranged according to a specially designed system of classification, the Universal Decimal Classification (UDC), has assembled a card catalog of many millions of entries; but because of the immense bulk of the material and the growing cost of compilation, it is unlikely that it will ever reach the goal set by its founders. The hopes for universal bibliographies have been largely replaced by the published catalogs of such great comprehensive libraries as the British Museum (1961–67), the Bibliothèque Nationale (beginning in 1897), the Library of Congress, and, most promising of all, the United States National Union Catalogue, maintained in the Library of Congress and, in the early 1970's, in the course of being printed in an estimated 625 volumes. Investigations were (early 1970's) being made into the problems involved in programming and computerizing such catalogs, and if the complexities of programming multilanguage material can be solved, if the necessary financing can be found, and if world cooperation in the production and standardization of catalog entries can be assured, it seems possible that the ideals of the Brussels Institute may yet to some extent be realized.

Hopes for a universal bibliography

Gesner's achievement was remarkable; besides writing many other books, including a number of bibliographical works, his universal bibliography with its appendix (1555) describes and classifies some 15,000 works by about 3,000 writers and provides alphabetical and subject indexes. His contemporaries turned their attention to detailed bibliographies of writings on particular subjects or in particular languages or relating to particular countries or places. Thus, John Bale, bishop of Ossory, published the earliest national bibliography, *Illustrium maioris Britanniae scriptorum . . . summarium* ("Summary of the Writings of the Most Eminent Britons," 1548), which lists British writers chronologically and sets out their writings in detail. The vastly improved extended second edition of this work, *Scriptorum illustrium maioris Brytanniae . . . catalogus* (1557), acknowledges Bale's indebtedness to John Leland, Henry VIII's library keeper and antiquary, who had made an antiquarian tour through England in 1534–42 and had made a survey of the writings of British authors. Other national bibliographies followed, such as those of the Italian Antonio Francesco Doni, *La Libraria del Doni Fiorentino Nella quale sono scritti tutti gl'autori vulgari* ("Library of . . . All Secular Authors," 1550: 2nd ed. also 1550); *La seconda del Doni* (1551); the Dutch theologian Cornelius Loos, *Illustrium Germaniae scriptorum catalogus* (1582); the French bibliographer François Grudé de la Croix du Maine, *Premier Volume de la bibliothèque du Sieur de la Croix du Maine* ("The First Volume of the Library of . . . ," 1584). The catalogs of the German book fairs held in Frankfurt and Leipzig each spring and autumn, beginning in 1564 and continuing until 1749, while not bibliographies in the strict sense, were nevertheless widely used as foundation material by early German bibliographers.

Catalogs of the German book fairs

Types of descriptive bibliographies. Most countries now have national bibliographies, in a majority of cases pub-

lished officially by the national library and based on copies of national publications deposited in accordance with provisions of copyright acts. Some notable exceptions are the United States and The Netherlands, whose national bibliographies are published commercially, and the United Kingdom, whose *British National Bibliography* is published by a council representing libraries, publishers, and booksellers. These national bibliographies aim at, and attain, a high degree of completeness and promptitude. The *British National Bibliography* (beginning in 1950), for example, is published weekly and cumulated quarterly and annually; it is arranged in a classified order according to the decimal classification of the American librarian Melvil Dewey, with an alphabetical index of authors, titles, and subjects. The *Bibliographie de la France* (beginning in 1811), published weekly, is arranged in a classified order with an annual index of authors, titles, and subjects. The German *Deutsche Bibliographie* (beginning in 1947) is published weekly and provides both an author and catchword index. The *Deutsche Bibliographie* has been published by computer since the beginning of 1968—the first such work to be so produced.

Bibliographies of books published in particular countries are of great value to students. Outstanding examples of such bibliographies are Charles Evans, *American Bibliography* (1903–34), covering the period 1639–1799; *A Short-Title Catalogue of Books Printed in England, Scotland and Ireland, and of English Books Printed Abroad, 1475–1640*, compiled by A.W. Pollard and G.R. Redgrave (1926; revised and enlarged 2nd ed., 1976), and its continuation by D.G. Wing, *Short-Title Catalogue . . . 1641–1700* (1945–51; revised 2nd ed., 2 vol., 1972–82). While not strictly bibliographies, the British Museum catalogs of its holdings of pre-1600 books from several European countries are similarly valuable because of the richness of the museum's collections.

Personal bibliographies may consist of no more than a simple list of an author's works, as, for example, those attached to articles in the *Dictionary of National Biography*. However, they may be more elaborate; take, for instance, F.W. Ebisch and L.L. Schücking, *A Shakespeare Bibliography* (1931; supplement 1937); Michael Sadleir, *Trollope: A Bibliography* (1928); Bertha Coolidge Slade, *Maria Edgeworth* (1937). Personal bibliographies are sometimes based on private collections, as exemplified by *A Stevenson Library* (1951–64), based on the collection of Edwin J. Beinecke; T.J. Wise's bibliographies, based on his own collections, of Tennyson, Swinburne, and others. A variant of personal bibliography consists of a narrative setting out an author's works with a comprehensive account of the circumstances surrounding the composition and publication of each work. An example of such a bibliography is J.E. Norton, *A Bibliography of the Works of Edward Gibbon* (1940).

Subject bibliographies vary in size, scope, and method, according to the purpose they are designed to serve. The following are examples of bibliographies that aim at offering comprehensive guidance: A. and A. de Backer and A. Carayon, *Bibliothèque de la Compagnie de Jésus*, ("Library of the Society of Jesus," 9 vol.; new ed., 1890–1909); F.W. Bateson, *The Cambridge Bibliography of English Literature* (4 vol., 1940; new ed. 1970 ff.); *Bibliography of British History*: C. Read, *Tudor Period, 1485–1603* (1933), G. Davies, *Stuart Period, 1603–1714* (1928), S. Pargellis and D.J. Medley, *The Eighteenth Century, 1714–1789* (1951); P. Caron, *Manuel pratique pour l'étude de la Révolution française* ("Practical Manual for the Study of the French Revolution"; latest ed. 1947); F.C. Dahlmann and G. Waitz, *Quellenkunde der deutschen Geschichte*. ("Published Sources of German History"; 9th ed. 1931–32); W.W. Greg, *A Bibliography of the English Printed Drama to the Restoration* (4 vol., 1939–59); G. Lanson, *Manuel bibliographique de la littérature française moderne* ("Bibliographic Manual of Modern French Literature"; latest ed. 1947); F. Madan, *The Early Oxford Press* (3 vol., 1895–1931); L.-N. Malclès, *Les Sources du travail bibliographique* ("Sources of Bibliographic Work." 3 vol., 1950–52). Bibliographies of publications on particular subjects, frequently including articles on periodicals, are numerous and cover many branches of human knowledge and interests.

Bibliographies of bibliographies

The proliferation of bibliographies has led to the publication of bibliographical guides—bibliographies of bibliographies. Among these should be mentioned *Bibliotheca Bibliographica* (1866), compiled by Julius Petzholdt, librarian of the Royal Library at Dresden, Germany, and *World Bibliography of Bibliographies* (5 vol.; 1965–66) by Theodore Besterman. Guides to current publications in a number of subject fields are published annually sometimes in narrative form—*e.g., The Year's Work in English-Studies* (1921 ff.) and *The Year's Work in Librarianship* (1929 ff.; later title *Five Years' Work in Librarianship*, 1958 ff.)—and sometimes in list form—*e.g., Internationale Bibliographie des Buch- und Bibliothekswesens* ("International Bibliography of Book and Library Systems," 1926–41).

Methods of compilation. The method of compiling a bibliography and the amount of detail included vary according to the author's purpose, his view of the importance of the subject, and his knowledge of it. The task of compiling a bibliography of any subject is a matter for a specialist in that subject. After an author has determined his specific objective and what material exists, he must proceed to describe and arrange it as usefully for his purpose as he can. In order to ensure that the descriptions are accurate and consistent, they should be made from the works themselves. If, for any reason, this proves impossible, the source of the information given should always be stated.

For most bibliographies it is usually necessary to give only the author, short title, place of imprint, and date; however, this information should be transcribed accurately from the book or article. Notes may be added on the scope and quality of the text and the location of a copy. In an author bibliography a chronological arrangement is frequently adopted, enabling the user to follow the development of the author's work. In such cases, later editions are listed with the earliest edition but additional mention is made at the appropriate chronological point. Subject bibliographies may be arranged in a systematic order to bring out the features important for the author's purpose, or a recognized classification system, such as the Dewey Decimal Classification or Universal Decimal Classification, may be employed.

Full descriptions are often given for early or rare books and for detailed author bibliographies. Elaborate rules have been evolved for compiling such descriptions, which make it possible for a skilled bibliographer to reconstruct from the text before him the makeup and appearance of a book. Such descriptions include information about the details of publication, number of copies printed, price, and binding. Semi-facsimile transcriptions are sometimes given, illustrating the various types of font used and the spacing of the title page. Modern methods of reproduction, however, make it possible to reproduce title pages photographically and thus obviate the use of such expedients, which in any case are not entirely satisfactory. The use of reproductions, however, has certain disadvantages, such as high cost and difficulty of layout, and, more important, reproductions themselves may be misleading, especially when the quality of the printing and paper of the original are poor. In 17th-century printing, for example, it is sometimes difficult to distinguish a mark of punctuation from a flaw in the paper, and a reproduction may give a wrong impression. If facsimiles are used, they should be as carefully edited as the text and the source of the reproduction should be stated.

Use of reproductions

The advent of computers and the application of data-processing techniques to library procedures such as cataloging have great potential value in bibliography. If a catalog is maintained in machine-readable form, it can be printed out in a number of different arrangements, or a variety of sublistings can be produced, with a relatively small increase in costs over the cost of printing the original arrangement. This method depends on the planning of requirements in advance and on the work of the programmer. It is thus possible for specialized lists, taken from the contents of a particular library, to be made available to a wide range of users. The best known and most highly

sophisticated retrieval project of this kind is the Medlars project (Medical Literature Analysis and Retrieval System), which has as its main objective the publication of the *Index Medicus* (a monthly listing of current articles from about 2,300 biomedical journals throughout the world) but which also supplies printed listings on particular subjects in the field of medicine.

The full benefit of these methods will accrue when it is economically and administratively possible to contemplate the storage of the enormous amounts of information available in the large comprehensive libraries.

In 1968 the Library of Congress began the production and distribution of cataloging data in machine-readable form. The scope of this enterprise was limited initially to monographic works in the English language, but with the intention of expanding the coverage to include other languages. The creation of the machine record of material already in the library before 1969, which is necessary if a full system of machine-readable records is to be created, was being investigated both in the Library of Congress, and elsewhere in other large comprehensive libraries. The problems likely to cause most difficulty are the labour involved, the great cost of the operation, and the magnitude of the data storage. Clearly such costly and complex operations have to be related to a realistic assessment of the benefits to be derived therefrom. Such assessments are in keeping with an interest, current in the early 1970s, in research library management. The rapid rise in library costs has emphasized the need for libraries, particularly large research libraries, to study whether and how they can apply accepted management techniques to the fuller exploitation of the bibliographical riches of their own institutions.

New approach to English literature

Critical bibliography. A valuable approach to the study of English literature, particularly, though not exclusively, of the 17th and 18th centuries, was developed about the turn of the 20th century out of a concern with Shakespearean texts in general and in particular with the nature and authority of the so-called quarto texts and their relation to the text printed in the First Folio. Editors had previously tended to assume that all the early texts of Shakespeare had equal authority and that it was perfectly proper to select readings which, in their view, improved the text of the author. The earlier editors were handicapped, in addition, by their ignorance of printing and book production in Shakespeare's time and by a lack of understanding of the nature of the texts they were dealing with. The change that took place stemmed from a variety of causes, of which the most important were probably two: (1) the close association in studies of English literature, especially drama, of three eminent scholars, A.W. Pollard, R.B. McKerrow, and W.W. Greg, all of them interested in bibliography; and (2) the remarkable progress made by Robert Proctor, a colleague of Pollard's in the British Museum, in the study of incunabula, the name given to books printed in the first 50 years after the invention of printing (*i.e.*, before 1501). The earliest printed books display the very individual characteristics of their printers in the type used, the typesetting, and the layout of the page, and it was found that, by studying these characteristics, useful deductions could be made about the place of individual books in a printer's total production and hence their approximate dates; in cases in which books contained no indication of the printer, it was often possible, using the same means, to assign a book to a particular printer. Important stages in the study of these earliest printed books, which had always been prized for their early date, their great aesthetic quality, and their rarity, were marked by a study of William Caxton, England's earliest printer, by William Blades, *The Life and Typography of William Caxton* (1861–63), in which typographical details were used to arrange Caxton's publications in chronological order; and by the work of a great antiquarian scholar and librarian, Henry Bradshaw, who made a special study of 15th-century books printed in the Netherlands. Bradshaw laid down an important principle: let the book speak for itself.

This new method was still more widely applied by Robert Proctor, who succeeded in allotting the large collections of incunabula in the British Museum and the Bodleian Library at Oxford to places of origin, by countries and towns, and to printers. Proctor's work firmly established this method of study, and the definitive catalog of the incunabula in the British Museum, which he began, demonstrated the valuable results that can be obtained by the meticulous examination of all the features of a book—paper, type, makeup, ornamentation, sewing, binding, manuscript notes, and marks of ownership.

It was against this background and, applying these methods to the literary and dramatic work of English authors of the 16th and 17th centuries, that Greg, McKerrow, and Pollard developed what has since come to be called critical, or analytical, bibliography, a method of bibliography that has been succinctly described as the study of books as tangible objects. By using the evidence of the physical features of books, it was often found possible to arrive at conclusions about the place of a book in an author's production that are beyond the scope of literary judgment.

Examination of the physical features of a book

One of the earliest and most illuminating examples of the application of this method concerns the question of the priority between the two issues of Shakespeare's play *Troilus and Cressida* that both bear a date of 1609. These differ in the preliminary matter: the title of one describes the play as having been acted at the Globe Theatre; in the other, the title page makes no mention of a production of the play, and an epistle to the reader stresses this fact by stating that the piece had never been "clapper-clawed with the palmes of the vulger." Literary considerations would suggest that the latter issue was the earlier. An examination of the books themselves, however, shows that the title page of the former is in its original state, while the latter's original title page had been cut away and two leaves substituted, containing the new title page and the address to the reader, thus demonstrating, without question, the order of the two.

In his edition of Beaumont and Fletcher's *Elder Brother* (1905), Greg similarly provided an irrefutable demonstration of the correct chronological order of the two quarto editions of 1637. In Q1 an improperly adjusted quad or space lead had produced a mark above the line before the word "young" in Act I, scene 2, line 72, which was mistaken by the compositor printing Q2 for an apostrophe and printed as such. As a consequence, Q2 was shown to have been printed from Q1 and not from the author's manuscript.

The success of such demonstrations led to a much more detailed study of the physical features of books and the bibliographical deductions that can be made from them. The methods of critical bibliography have been substantially developed and have been widely applied to books of later periods, and it has been shown that, even in the sophisticated machine age, careful bibliographical examination has a significant role to play in determining the reliability of an author's text and the place of a book in an author's production.

Exposure of false first editions

The method was spectacularly displayed in the exposure of a number of alleged first editions of poems, essays, and other minor productions of well-known 19th-century authors, such as the "Reading" sonnets of Elizabeth Barrett Browning—*i.e.*, an edition of these sonnets, dated 1847, with Reading as the place of imprint. These books, more than 50 in number, were shown, in *An Enquiry into the Nature of Certain Nineteenth Century Pamphlets* (1934), by John Carter and Graham Pollard, to have been printed on kinds of paper, made from esparto grass or wood pulp, known not to have been in use at the dates shown on the title pages, and using printing types that were first cut much later.

It has been clearly demonstrated that machine-printed books of the 19th and 20th centuries produce textual problems rivalling those found in hand-printed books. While editorial procedures and the principles governing the interpretation of bibliographical evidence remain the same, new printing methods force bibliographers to devise new procedures to deal with the new problems presented to them.

A widely used guide for the bibliographical study of

English books up to the end of the 17th century is provided by R.B. McKerrow's *Introduction to Bibliography for Literary Students* (1927). The publications of the Bibliographical Society of London, especially its journal, *The Library* (1889 ff.), contain valuable information about bibliographical methods and their application. A useful account of the work of the society and a study by F.P. Wilson of the application of bibliographical techniques to the works of William Shakespeare, "Shakespeare and the New Bibliography," is contained in *The Bibliographical Society, 1892–1942: Studies in Retrospect* (1945). Other bibliographical journals that may be usefully consulted are: *The Papers of the Bibliographical Society of America* (1904 ff.), the *Publications* of the *Oxford Bibliographical Society* (1922 ff.), and the *Transactions of the Cambridge Bibliographical Society* (1949 ff.). *Studies in Bibliography* (1948 ff.), published by the Bibliographical Society of the University of Virginia, is especially valuable. (F.C.F.)

CHRONOLOGY

Chronology, in the broadest sense, is a time scale, a method of ordering time. This article, however, deals mainly with systems of chronology used by different peoples in recording their history.

Scientific chronology, which seeks to place all happenings in the order in which they occurred and at correctly proportioned intervals on a fixed scale, is used in many disciplines and can be utilized to cover vast epochs. Astronomy, for example, measures the sequence of cosmic phenomena in thousands of millions of years; geology and paleontology, when tracing the evolution of the Earth and of life, use similar epochs of hundreds or thousands of millions of years. Geochronology reckons the more distant periods with which it deals on a similar scale; but it descends as far as human prehistoric and even historic times, and its shorter subdivisions consist only of thousands of years. Shortest of all are the chronological scales used in the recording of human events in a more or less systematic and permanent manner. These vary in scope, accuracy, and method according to the purpose, degree of sophistication, and skill of the peoples using them, as do the calendrical systems with which they are inextricably bound up. For further details see the article CALENDAR.

It is difficult to fix ancient historical chronologies in relation to scientific chronology. The terms of reference of ancient peoples were vague and inconsistent when judged by modern standards, and many of their inscriptions and writings have inevitably disappeared. The gaps in their records are increasingly filled in and their inconsistencies removed by the results of archaeological excavation. Guided by these findings, scholars can confirm, refute, or amend chronological reconstructions already tentatively made. Astronomical calculation and dating by radioactive-carbon content are also helpful in the work of fixing ancient chronologies.

Chinese. Chinese legendary history can be traced back to 2697 BC, the first year of Huang Ti (Chinese: Yellow Emperor), who was followed by many successors and by the three dynasties, the Hsia, the Shang, and the Chou. Recent archaeological findings, however, have established an authentic chronology beginning with the Shang dynasty, though the exact date of its end remains a controversial topic among experts. The so-called oracle-bone inscriptions of the last nine Shang kings (1324–1122 BC) record the number of months up to the 12th, with periodical additions of a 13th month, and regular religious services on the summer and winter solstice days, all of which indicates the adjustment of the length of the lunar year by means of calculations based on the solar year. Individual days in the inscriptions are named according to the designations in the sexagenary cycle formed by the combination of the 10 celestial stems and 12 terrestrial branches. Every set of 60 days is divided into six 10-day "weeks." Also recorded are numerous eclipses that can be used to verify the accuracy of the Shang chronology. In the oracular sentences of the last Shang king, Chou Hsin, the year of his reign is referred to as "the King's *n*th annual sacrifice."

From the beginning of the following (Chou) dynasty, the word year was etymologically identical with "harvest." Thus, "King X's *n*th harvest" meant the *n*th year of his reign. The lunar month was then divided into four quarters—Ch'u-chi, Tsai-sheng pa, Chi-sheng pa and Chi-szu pa—and the practice of using the 60 cyclical names for the days was continued. Thus, in the inscription on a Chou bronze vessel, a typical date would read: "In the King's *n*th harvest, in the *n*th quarter of the *n*th month, on the day X-*y*, etc."

The tradition of recording events by referring to the king's regnal year continued until 163 BC, when a new system, *nien-hao* ("reign-period title"), was introduced by Emperor Han Wen Ti of the Former Han dynasty (206 BC–AD 8). Thereafter, every emperor proclaimed a new *nien-hao* for his reign at the beginning of the year following his accession (sometimes an emperor redesignated his *nien-hao* on special occasions during his reign). A typical date in the *nien-hao* system might read, "the third year of the Wan-li reign period" (Wan-li san nien). In order to date any event in Chinese history, it is necessary to convert the year in the period of the designated *nien-hao* into the Western calendar.

The *nien-hao* system

During the Chou dynasty the civil year began with the new moon, which occurred before or on the day of the winter solstice. This "first month" of the Chou year (Chou *cheng*) was equivalent to the 11th month of the Hsia year (Hsia *cheng*) or to the 12th month of the Shang year. The first emperor, Shih Huang-ti, of the short-lived Ch'in dynasty (221–206 BC) made the year begin one month earlier—*i.e.*, with the lunation (the period of time between one new moon and the next) before the one in which the winter solstice occurred. The Ch'in year was continuously used until 104 BC, when Emperor Han Wu Ti promulgated the T'ai-ch'u calendar by reverting to the Hsia *cheng*—*i.e.*, by taking the third month of the Chou year, or the second lunation after the winter solstice, as the first month of the civil year. This lunar year (or Hsia *cheng*) was used till the last day of the Ch'ing, or Manchu, dynasty (1644–1911/12). When in 1911 the first republic was founded, the solar year was officially adopted, but successive governments kept the *nien-hao* tradition by referring any date to the number of years since the establishment of the republic—*e.g.*, 1948 was chronicled "the 37th year of the republic." In 1949, when the People's Republic of China was proclaimed, the old system was replaced by the Gregorian calendar.

(W.S.-cg.)

Japanese. The principal chronicles describing the origins of Japanese history are the *Nihon shoki* ("Chronicle of Japan") and the *Koji-ki* ("Record of Ancient Matters"). The *Nihon shoki* (compiled in AD 720) assembled information in a chronological order of days, months, and years starting several years before 660 BC, which was the year of the enthronement of the first Japanese emperor, who was posthumously named Jimmu. The *Koji-ki* (compiled in AD 712) related events under the reign of each emperor without a strict chronological order. Sometimes the *Koji-ki* gave the years of emperors' deaths and their ages at death. This information is different from that recorded in the *Nihon shoki*.

Native Japanese scholars since Fujiwara Teikan in the 18th century have realized that the *Nihon shoki* was historically inadequate and different from the *Koji-ki*, at least insofar as the chronological information is concerned. They have suggested that the foundation year of Japan was 600 years later than stated in the *Nihon shoki*. Naka Michiyo (late 19th century) argued with minute detail about the question of Japanese chronology. His ideas were supplemented by those of other Japanese scholars, who pointed out that: (1) the reigns of the earlier Japanese emperors as stated in the *Nihon shoki* are unnaturally long; (2) the date of the enthronement of the emperor Jimmu should be reconsidered; (3) a chronological gap exists between the *Nihon shoki* and contemporary Chinese and Korean chronicles. In comparison with Korean chronicles, they argued, the *Nihon shoki* has created an intentional expansion of chronology—*i.e.*, the entries about the empress Jingō and the emperor Ōjin can be identified with historical facts relating to the Korea of the 4th and 5th centuries and therefore must be placed 120 years later

Reforms of Naka Michiyo

than mentioned in the *Nihon shoki*. When comparing the *Nihon shoki* with Chinese chronicles, one finds the chronological gap somewhat reduced. The Chinese chronicles provide information about the tributes sent individually by five Japanese "kings" to Liu-Sung and Southern Ch'i during the 5th century. There are still questions about the identification of these kings, but it is generally accepted that the "king" written in Chinese character as Wu must be the Japanese emperor Yūryaku. By the late 5th century the gap between Japanese and Korean records, on the one hand, and Japanese and Chinese, on the other hand, disappears.

The intentional expansion of the chronology of the *Nihon shoki* was adopted by its compilers, who identified Queen Himiko (Pimihu) of Yamatai of the chronicle of Wei China with the Empress Jingō of Japanese legend.

The method of designating a year by the *kan-shi* (sexagenary cycle) appears to have begun about the reign of Emperor Yūryaku, when, as mentioned above, the gap between the continental and Japanese chronologies was bridged. The inscription on remarkable copper images of Buddha cast just after the period of Prince Shōtoku's regency (AD 593–621) bears a *nengō* (*nien-hao*, or reign-year title), although not a strictly authorized one. It was at this time that the Chinese luni-solar calendar system was adopted. The first official *nengō* was Taika, which was adopted by the imperial court in 645. Since 701, when the second title, Taihō, was adopted, the reign-year system has been continuously used in relation to the emperors' reigns up to the present day. In medieval times Japanese chronology underwent a remarkable evolution: (1) when the Imperial dynasty split into two courts (1336–92), two series of *nengō* began to be used; (2) during the Ashikaga period some private *nengō* again appeared; (3) some dates of the authorized "central" calendars did not correspond with those of locally compiled calendars. Moreover, military leaders would not accept some of the new *nengō*. Minamoto Yoritomo, for example, did not use the *nengō* that was adopted by the emperor Antoku and the Taira regime, and Ashikaga Mochiuji and Ashikaga Shigeuji did not use the official, respectively Eikyō and Kōshō, *nengō*.

In the Tokugawa period (1603–1867), gaps between central and provincial calendars disappeared, especially after the establishment of the Jōkyō calendar, the first native calendar compiled in Japan, instead of the Chinese-based one that was in use until this period. On January 1, 1873, Emperor Meiji adopted the Gregorian calendar in use in the West and at the same time adopted the "Japanese Era," with Emperor Jimmu as its founder, in addition to the *nengō* system. (Hi.Mo.)

Indian. Two kinds of chronological systems have been used in India by the Hindus from antiquity. The first requires the years to be reckoned from some historical event. The second starts the reckoning from the position of some heavenly body. The historical system, the more common in modern times, exists side-by-side with Muslim and international systems successively introduced.

Reckonings dated from a historical event. The inscriptions of the Buddhist king Aśoka (*c.* 265–238 BC) give the first epigraphical evidence of the mode of reckoning from a king's consecration (*abhiṣeka*). In these inscriptions (Middle Indian language in India or Greek and Aramaean in what is now Qandahār, Afghanistan) the dates are indicated by the number of complete years elapsed since the king's consecration. But the earlier existence of a reckoning of duration of reigns and dynasties is evidenced by the testimony of the Greek historian Megasthenes, who in 302 BC was the ambassador of Seleucus I Nicator, founder of the Seleucid Empire, to the court of Candragupta Maurya, Aśoka's grandfather. According to Megasthenes, the people of the Magadha kingdom, with its capital Pāṭaliputra (Patna), kept very long dynastic lists, preserved in the later Sanskrit Purāṇas (legends of the gods and heroes) and later Buddhist and Jain chronicles. They generally indicate, in years or parts of years, the duration of each reign.

Similar records of other periods and regions exist, and a relative chronology may be established. Unfortunately, it is not always possible to connect them with any absolute chronology, the precise dates of the reigns given being still unsettled. For example, in the Scythian period of the history of northern India, several inscriptions are dated from the beginning of the reign of Kaniṣka, the greatest king of the Asian (Kushān) invaders, but his dates are still uncertain (AD 78, 128–129, 144, etc., have been suggested for the beginning of a Kaniṣka era).

Other records give regnal years that can be linked with absolute chronology through other data—*e.g.*, those of several rulers of the Rāṣṭrakūṭa of the Deccan.

Hindu dynastic eras

The dynastic eras, founded by several rulers and kept up or adopted by others, are also numerous. The most important were the Licchavi era (AD 110), used in ancient Nepal; the Kalacuri era (AD 248), founded by the Abhūri king Īśvarasena and first used in Gujarāt and Mahārāshtra and later (until the 13th century) in Madhya Pradesh and as far north as Uttar Pradesh; the Valabhi era (AD 318, employed in Saurāṣṭra) and the Gupta era (AD 320), used throughout the Gupta Empire and preserved in Nepal until the 13th century. Later came the era of the Thakuri dynasty of Nepal (AD 395), founded by Aṃśuvṛrman; the Harṣa era (AD 606), founded by Harṣa (Harṣavardhana), long preserved also in Nepal; the western Calukya era (AD 1075), founded by Vikramāditya VI and fallen into disuse after 1162; the Lakṣmaṇa era (AD 1119), wrongly said to have been founded by the king Lakṣmaṇasena of Bengal and still used throughout Bengal in the 16th century and preserved until modern times in Mithilā; the Rājyabhiṣekasaka or Marāthā era (1674), founded by Śivāji but ephemeral.

Later, instead of the beginning of a reign or of a dynasty, the death of a religious founder was adopted as the starting point of an era. Among Buddhists the death of the Buddha and among the Jains the death of the Jina were taken as the beginning of eras. The Jain era (*vīrasaṃvat*) began in 528 BC. Several Buddhist sects (no longer existing in India) adopted different dates for the death (Nirvāṇa) of the Buddha. The Buddhist era prevailing in Ceylon and Buddhist Southeast Asia begins in 544 BC.

Historical events, now obscure, were the basis of the two most popular Indian eras: the Vikrama and the Śaka.

The Vikrama era (58 BC) is said in the Jain book *Kālakācāryakathā* to have been founded after a victory of King Vikramāditya over the Śaka. But some scholars credit the Scytho-Parthian ruler Azes with the foundation of this era. It is sometimes called the Mālava era because Vikramāditya ruled over the Mālava country, but it was not confined to this region, being widespread throughout India. The years reckoned in this era are generally indicated with the word *vikramasaṃvat*, or simply *saṃvat*. They are elapsed years. In the north the custom is to begin each year with Caitra (March–April) and each month with the full moon. But in the south and in Gujarāt the years begin with Kārttika (October–November) and the months with the new moon; in part of Gujarāt, the new moon of Āṣāḍha (June–July) is taken as the beginning of the year. To reduce Vikrama dates to dates AD, 57 must be subtracted from the former for dates before January 1 and 56 for dates after.

The Śaka era

The Śaka, or Śālivāhana, era (AD 78), now used throughout India, is the most important of all. It has been used not only in many Indian inscriptions but also in ancient Sanskrit inscriptions in Indochina and Indonesia. The reformed calendar promulgated by the Indian government from 1957 is reckoned by this era. It is variously alleged to have been founded by King Kaniṣka or by the Hindu king Śālivāhana or by the satrap Nahapāna. According to different practices, the reckoning used to refer to elapsed years in the north or current years in the south and was either solar or luni-solar. The luni-solar months begin with full moon in the north and with new moon in the south. To reduce Śaka dates (elapsed years) to dates AD, 78 must be added for a date within the period ending with the day equivalent to December 31 and 79 for a later date. For Śaka current years the numbers to be added are 77 and 78. The official Śaka year is the elapsed year, starting from the day following that of the vernal equinox. A normal year consists of 365 days, while the leap year has 366. The first month is Chaitra, with 30 days in a normal year and

31 in a leap year; the five following months have 31 days, the others 30.

A Nepalese era (AD 878) of obscure origin was commonly used in Nepal until modern times. The years were elapsed, starting from Kārttika, with months beginning at new moon. Another era, the use of which is limited to the Malabār Coast (Malayalam-speaking area) and to the Tirunelveli district of the Tamil-speaking area, is connected with the legend of the hero Paraśurāma, an avatar (incarnation) of the god Vishnu. It is called the Kollam era (AD 825). Its years are current and solar; they start when the Sun enters into the zodiacal sign of Virgo in north Malabār and when it enters into Leo in south Malabār. It is sometimes divided into cycles of 1,000 years reckoned from 1176 BC. Thus, AD 825 would have been the first year of the era's third millennium.

Eras based on astronomical speculation. During the period of elaboration of the classical Hindu astronomy, which was definitively expounded in the treatises called *siddhāntas* and by authors such as Āryabhaṭa I (born AD 476), Varāhamihira, Brahmagupta (7th century AD), etc., the ancient Vedic notions on the cycle of years, embracing round numbers of solar and lunar years together, were developed. On the one hand, greater cycles were calculated in order to include the revolutions of planets, and the theory was elaborated of a general conjunction of heavenly bodies at 0° longitude after the completion of each cycle. On the other hand, cosmologists speculated as to the existence of several successive cycles constituting successive periods of evolution and involution of the universe. The period calculated as the basis of the chronology of the universe was the *mahāyuga,* consisting of 4,320,000 sidereal years. It was divided into four yugas, or stages, on the hypothesis of an original "order" (*dharma*) established in the first stage, the Kṛta Yuga, gradually decaying in the three others, the Tretā, Dvāpara, and Kali yugas. The respective durations of these four yugas were 1,728,000, 1,296,000, 864,000, and 432,000 years. According to the astronomer Āryabhaṭa, however, the duration of each of the four yugas was the same—*i.e.,* 1,080,000 years. The basic figures in these calculations were derived from the Brahmanical reckoning of a year of 10,800 *muhūrta* (see CALENDAR: *The Hindu calendar*), together with combinations of other basic numbers, such as four phases, 27 *nakṣatra*s, etc. The movement of the equinoxes was at the same time interpreted not as a circular precession but as a libration (periodic oscillation) at the rate of 54 seconds of arc per year. It is in accordance with these principles that the calculation of the beginning of the Kali Yuga was done in order to fix for this chronology a point starting at the beginning of the agreed world cycle. Such a beginning could not be observed, since it was purely theoretical, consisting of a general conjunction of planets at longitude 0°, the last point of the *nakṣatra* Revatī (Pisces). It has been calculated as corresponding to February 18, 3102 BC (old style), 0 hour, and taken as the beginning of the Kali era. In this era, the years are mostly reckoned as elapsed and solar or luni-solar.

Cosmic cycles

In Hindu tradition the beginning of the Kali era was connected with (1) events of the Mahābhārata war; (2) King Yudhiṣṭhira's accession to the throne; (3) 36 years later, King Parikṣit's consecration; and (4) the death of Lord Krishna. Years of the era are still regularly given in Hindu almanacs.

An era resting upon a fictitious assumption of a complete 100-year revolution of the Ursa Major, the Great Bear (*saptarṣi*), around the northern pole was the Saptarṣi, or Laukika, era (3076 BC), formerly used in Kashmir and the Punjab. The alleged movement of this constellation has been used in Purāṇa compilations and even by astronomers for indicating the centuries.

Two chronological cycles were worked out on a basis of the planet Jupiter's revolutions, one corresponding to a single year of Jupiter consisting of 12 solar years and the other to five of Jupiter's years. The second, the *bṛhaspaticakra,* starts, according to different traditions, from AD 427 or from 3116 BC. Before AD 907 one year was periodically omitted in order to keep the cycle in concordance with the solar years. Since 907 the special names by which every year of the cycle is designated are simply given to present years of the almanac.

Side-by-side with Hindu and foreign eras adopted in India, several eras were created in the country under foreign influence, chiefly of the Mughal emperor Akbar: Bengali San (AD 593), Amli of Orissa and Vilayati (AD 592), Faṣlī (AD 590, 592, or 593 according to the district), and Sursan of Mahārāshtra (599). (J.L.A.F.)

Egyptian. At the end of the 4th millennium BC, when King Menes, the first king of a united Egypt, started his reign, the ancient Egyptians began to name each year by its main events, presumably to facilitate the dating of documents. These names were entered into an official register together with the height of the Nile during its annual inundation. Short notes at first, the year names developed into lengthy records of historical and religious events, especially of royal grants to the gods. These lists grew into annals, which were kept during the entire history of Egypt so that later kings could, after important events, consult the annals and ascertain whether a comparable occurrence had happened before. Unfortunately, these annals are lost. Only fragments from the 1st to the 5th dynasty (*c.* 3100–*c.* 2345 BC) are preserved, copied on stone. These fragments, however, are in such poor condition that they raise more chronological problems than they solve.

The Egyptian priests of the Ramesside period (*c.* 1300 BC) copied the names and reigns of the kings from Menes down to their time from the annals, omitting all references to events. Even this king list would have given a safe foundation of an Egyptian chronology, but the only extant copy, on a papyrus now kept at the Museo Egizio in Turin, has survived only in shreds, entire sections having been lost. Extracts from this king list, which name only the more important kings, are preserved in the temples of the kings Seti I and Ramses II at Abydos and on the wall of a private tomb at Ṣaqqārah (now in the Egyptian Museum), but they give little help in chronological matters.

The Turin Papyrus

When the Greeks began to rule Egypt after the conquest of Alexander the Great, King Ptolemy II Philadelphus, hoping to acquaint the new ruling class with the history of the conquered country, commissioned Manetho, an Egyptian priest from Sebennytus, to write a history of Egypt in the Greek language. As Manetho had access to the ancient annals, he added some of their entries to his list of kings and reigns, especially during the first dynasties. The more he progressed in time, the more he added semihistorical traditions and stories as they were composed by the Egyptian priests to discuss moral problems in the disguise of a historical "novel." There had been, undoubtedly, fewer historical facts in Manetho's history than one might expect. But Manetho's work, too, is lost except for some excerpts used by Sextus Julius Africanus and Eusebius in writing their chronicles. These, in turn, represented the material used in part by George Syncellus in the 8th century AD. During copying and recopying, Manetho's text clearly suffered many changes, unintentionally or on purpose. The figures of the reigns, especially of the older dynasties, for instance, were enlarged when some of the early Christian historians tried to equate King Menes with Adam. In addition, the excerpts were done carelessly. Therefore, Manetho's work, as handed down to us, is short of useless. Nevertheless, together with the fragments of the annals and of the king list of Turin, they create a framework of Egyptian chronology; so the division into dynasties was taken over from Manetho. But to achieve a continuous history of Egypt and to bridge the gaps left by the fragmentary state of the extant chronological material, scholars must turn to other means, particularly astronomical references found in dated texts. These are related principally to the rising of Sothis and to the new moon.

Theoretically, the Egyptian civil year began when the Dog Star, Sirius (Egyptian Sothis), could first be seen on the eastern horizon just before the rising of the Sun (*i.e.,* 19/20 of July). As the civil calendar of the ancient Egyptians consisted of 12 months (each of 30 days) and five odd days (called epagomenal days), the civil year was a quarter of a day too short in relation to the rising of Sothis, so that the new year advanced by one day every four years. New Year's Day and the rising of Sothis coincided again

The Sothic cycle

only after approximately 1,460 years, the so-called Sothic cycle. Dated documents mentioning the rising of Sothis can be translated into the present calendar by multiplying the number of days elapsed since the first day of the year by four and subtracting this sum from the date of the beginning of the particular Sothic cycle. The dates for the start of each Sothic cycle are fortunately known because the Roman historian Censorinus fixed the coincidence of New Year's Day and heliacal rising of Sothis in AD 139. Taking into account a slight difference between a Sothic year and a year of the fixed stars, the years 1322, 2782, and 4242 BC are taken as starting points of a Sothic cycle.

There are six ancient Egyptian documents extant giving Sothis dates, but only three of these are of value. The oldest is a letter from the town of Kahun warning a priest that the heliacal rising of Sothis will take place on the 16th day of the 8th month of year 7 of a king who, according to internal evidence, is Sesostris III of the 12th dynasty. This date corresponds to 1866 BC, according to the corrected Sothic cycle. The next date is given by a medical papyrus written at the beginning of the 18th dynasty, to which a calendar is added, possibly to ensure a correct conversion of dates used in the receipts to the actual timetable. Here it is said that the 9th day of the 11th month of year 9 of King Amenhotep I was the day of the heliacal rising of Sothis—*i.e.,* 1538 BC. This date, however, is only accurate provided that the astronomical observations were taken at the old residence of Memphis; if observed at Thebes in Upper Egypt, the residence of the 18th dynasty, the date must be lowered by 20 years—*i.e.,* 1518 BC. The third Sothis date shows that Sirius rose heliacally sometime during the reign of Thutmose III, which lasted for 54 years, on the 28th day of the 11th month; so year 1458 BC (point of observation at Memphis) or 1438 BC (point of observation at Thebes) must have belonged to the reign of this king. From these dates it is possible to calculate the absolute dates for the reigns of the 12th dynasty, as the durations of most of the reigns of the kings belonging to this dynasty are preserved on the king list of the Turin Papyrus. On the other hand, chronologists are able to compute the reigns of the kings of the 18th dynasty by utilizing the highest dates of their documents and the figures preserved by Manetho. Historians are also helped by the fact that the Egyptians sometimes identified a certain day as "exactly new moon"; they reckoned new moon from the morning after the last crescent of the waning moon had become invisible in the east just before sunrise. As there is a 25-year lunar cycle, such ancient Egyptian moon dates could be calculated with a fair amount of certainty but of course only if the ancient Egyptians themselves observed this celestial phenomenon accurately. There is some doubt, however, as it is shown by the attempts of very competent scholars to convert these moon dates. Sometimes even moon dates given by the same papyrus contradict themselves; in another case, the date given by a document had to be amended to achieve a reasonable result. These and other examples show that ancient Egyptian statements on celestial phenomena, especially on new moons, tend to be inaccurate because of faulty or inexact observations. Therefore, every date given for a fixed reign should be used with caution as the astronomical observation on which it is based may be inexact. Sometimes they are controlled by synchronism with Babylonian, Assyrian, or Hittite king lists or, later on, by the close interconnections between Greek and Egyptian history. Sometimes even biographical data are helpful. The statements found on small stelae inside the burial ground of the holy bulls of Memphis (Apis) register the dates of birth, enthronement, and death of these animals accurately. But the more time recedes, the more the chronology of the Egyptian history becomes uncertain, even when astronomical data are available. Up till now even carbon-14 data are of no great help, as uncertainties are mostly not greater than the standard deviations to be expected in a carbon-14 calculation.

Nevertheless, Egyptologists believe themselves to be on fairly firm ground when dating the beginning of the Ancient Kingdom (1st and 2nd dynasty) about 3090 BC, the beginning of the 11th dynasty at 2133 BC, and of the Middle Kingdom (12th dynasty) at 1991 BC. The New Kingdom started at 1567 or 1552 BC, depending on a choice for the first year of Ramses II of either 1290 BC or 1304 BC—one lunar cycle earlier. The following centuries still pose many chronological questions down to 664 BC, when Greek historiography took over. (W.H.)

Babylonian and Assyrian. *Mesopotamian chronology, 747 to 539 BC.* The source from which the exploration of Mesopotamian chronology started is a text called Ptolemy's Canon. This king list covers a period of about 1,000 years, beginning with the kings of Babylon after the accession of Nabonassar in 747 BC. The text itself belongs to the period of the Roman Empire and was written by a Greek astronomer resident in Egypt. Proof of the fundamental correctness of Ptolemy's Canon has come from the ancient cuneiform tablets excavated in Mesopotamia, including some that refer to astronomical events, chiefly eclipses of the Moon. Thus, by the time excavations began, a fairly detailed picture of Babylonian chronology was already available for the period after 747 BC. Ptolemy's Canon covers the Persian and Seleucid periods of Mesopotamian history, but this section will deal only with the period up to the Persian conquest (539 BC).

The chief problem in the early years of Assyriology was to reconstruct a sequence for Assyria for the period after 747 BC. This was done chiefly by means of *limmu,* or eponym, lists, several of which were found by early excavators. These texts are lists of officials who held the office of *limmu* for one year only and whom historians also call by the Greek name of eponym. Annals of the Assyrian kings were being found at the same time as eponym lists, and a number of these annals, or the campaigns mentioned in them, were dated by eponyms who figured in the eponym lists. Moreover, some of the Assyrian kings in the annals were also kings of Babylonia and as such were included in Ptolemy's Canon.

The *limmu* lists

Good progress was therefore being made when, soon after 1880, two chronological texts of outstanding importance were discovered. One of these, now known as King List A, is damaged in parts, but the end of it, which is well preserved, coincides with the first part of Ptolemy's Canon down to 626 BC. The other text, *The Babylonian Chronicle,* also coincides with the beginning of the canon, though it breaks off earlier than King List A. With the publication of these texts, the first phase in the reconstruction of Mesopotamian chronology was over. For the period after 747 BC, there remained only one serious lacuna—*i.e.,* the lack of the eponym sequence for the last 40 years or so of Assyrian history. This had not been established by the early 1970s.

Assyrian chronology before 747 BC. German excavations at Ashur, ancient capital of Assyria, yielded further eponym lists. By World War I the full sequence of eponyms was known from about 900 to 650 BC. A further fragmentary list carried the record back to about 1100 BC, and on this basis Assyrian chronology was reconstructed, with little error, back to the first full regnal year of Tiglathpileser I in 1115 BC. Without another eponym list, a king list was needed for substantial further progress. King lists found at Ashur proved disappointing. Those fairly well preserved did not include figures for the reigns, and those with figures were very badly damaged.

In 1933, however, an expedition from the University of Chicago discovered at Khorsabad, site of ancient Dur Sharrukin, an Assyrian king list going back to about 1700 BC. But for the period before 1700 BC the list is damaged and otherwise deficient, and Assyrian chronology prior to this date is still far from clear.

Before 747 BC it was the custom of the Assyrian kings to hold eponym office in their first or second regnal year. Thus, in an eponym list, the number of names between the names of two successive kings usually equals the number of years in the reign of the first of the two kings. It would have been easy to compile a king list from an eponym list, and there is evidence that this Assyrian king list was compiled from an eponym list probably in the middle of the 11th century BC. As an eponym list is a reliable chronological source, since omission of a name entails an error of only one year, the king list, if based

on one, will have preserved much of the structure of older eponym lists now lost. (Except for one fragment, no known eponym list goes back further than the beginning of the 11th century BC.)

Babylonian chronology before 747 BC. In the long interval between the fall of the last Sumerian dynasty *c.* 2000 BC and 747 BC there are two substantial gaps in chronology, each about two centuries long. The earlier gap is in the 2nd millennium, from approximately 1600–1400 BC, the later gap in the 1st millennium, from *c.* 943–747 BC. During these gaps the names of most of the kings are known, as well as the order, but usually not the length of their reigns.

A means of checking the reliability of the Babylonian king list is provided by the chronicles, annals, and other historical texts that show that a given Assyrian king was contemporaneous with a given Babylonian king. There are no fewer than 15 such synchronisms between 1350 and 1050 BC, and, when the Babylonian and Assyrian king lists are compared, they all fit in easily. Only one of them, however, provides a close approximate date in Babylonian chronology. This synchronism shows that the two-year reign of the Assyrian king Ashared-apil-Ekur (*c.* 1076–*c.* 1075 BC) is entirely comprised within the 13-year reign of the Babylonian king Marduk-shapik-zeri. The Assyrian's dates are probably correct to within one year. Thus, if Marduk-shapik-zeri is dated so that equal proportions of his reign fall before and after that of Ashared-apil-Ekur, a date is obtained for the former that should not be in error more than six years. This synchronism constitutes a key to the structure of Babylonian chronology by providing the base date for all the reigns in the interval *c.* 1400–943 BC for which the Babylonian king list gives figures. All the dates thus obtained are subject to the six-year margin of error.

These synchronisms between Assyrian and Babylonian kings continue throughout the period that corresponds to the second gap in the Babylonian king list—from *c.* 943–747 BC. Since the Assyrian chronology in that period is firmly established, these synchronisms provide a useful framework for the structure of Babylonian chronology in that period.

The gap in the 2nd millennium BC, however, is not as easy to fill. The fact that the magnitude of the gap is uncertain constitutes the main problem in the chronology of the 2nd millennium BC and also affects the chronology of the preceding Sumerian period. The problem is not yet solved. Observations of the planet Venus made during the reign of King Ammisaduqa, less than 50 years before the end of the 1st dynasty of Babylon, permit only certain possible dates for his reign. Translated into dates for the end of the dynasty, the three most likely possibilities are 1651, 1595, and 1587 BC. The evidence is not yet conclusive and leaves uncertain what choice should be made among the three. The chronology adopted here is based on the second of these dates for the end of the 1st Babylonian dynasty—*i.e.*, 1595 BC.

Prior to this gap in the 2nd millennium BC, there is a period of five centuries with a well-established chronological structure. All the kings in the major city-states are known, as well as their sequence and the length of their reigns. Which sets of dates should be assigned to these reigns, however, depends on the date adopted for the 1st dynasty of Babylon. This period of five centuries extends from the beginning of the 3rd dynasty of Ur to the end of the 1st dynasty of Babylon—*i.e.*, on the chronology adopted here, 2113–1595 BC. During this period the Babylonians dated their history not by regnal years but by the names of the years. Each year had an individual name, usually from an important event that had taken place in the preceding year. The lists of these names, called year lists or date lists, constitute as reliable a source in Babylonian chronology as the eponym lists do in Assyrian chronology. One of the events which almost invariably gave a name to the following year was the accession of a new king. Hence, the first full regnal year of a king was called "the year (after) NN became king." In Assyria the number of personal names in an eponym list between the names of two successive kings normally equalled the number of years in the reign of the first king, and, similarly, in Babylonia the number of year names between two year names of the above kind nearly always equalled the number of years in the reign of the first king. Just as in Assyria, the eponym lists are almost certainly the source of the king lists, so in Babylonia the king lists are based on the year lists. Several of these king lists, compiled at a time when the year lists were still in use, survive. One gives the 3rd dynasty of Ur and the dynasty of Isin; another gives the dynasty of Larsa. Both may be school texts.

Babylonian year lists

The 3rd dynasty of Ur and the dynasty of Isin also figure in the Sumerian king list, which reaches far back into the Sumerian period. The original version probably ended before the 3rd dynasty of Ur, but later scribes brought it up to date by adding that dynasty as well as the dynasty of Isin.

(M.B.R.)

Jewish. The era at present in vogue among the Jews, counted from the creation of the world (*anno mundi*; abbreviated to AM), came into popular use about the 9th century AD. Traceable in dates recorded much earlier, this era has five styles conventionally indicated by Hebrew letters used as numerals and combined into mnemonics, which state the times of occurrence of the epochal mean conjunctions of *moladim* (see CALENDAR: *Middle Eastern calendar systems: The Jewish calendar*) or the orders of intercalation in the 19-year cycle or both. The respective epochs of these styles fall in the years 3762–3758 BC, inclusive. By about the 12th century AD the second of the mentioned styles, that which is in use at present, superseded the other styles of the era *anno mundi*.

The styles of this era arise from variations in the conventional rabbinical computation of the era of the creation. This computation, like hundreds of other calculations even more variable and no less arbitrary, is founded on synchronisms of chronological elements expressed in the terms of biblical and early postbiblical Jewish eras.

The biblical era *anno mundi* underlies the dating of events (mainly in the book of Genesis) prior to the Exodus from Egypt. This period of biblical chronology abounds in intractable problems caused by discrepancies between the Jewish and Samaritan Hebrew texts and the Greek version known as the Septuagint, by apparent inconsistencies in some of the synchronisms, and by uncertainties about the method of reckoning.

Discrepancies in biblical texts

During the period from the Exodus to the founding of Solomon's Temple, the only continuous biblical era (chiefly in the remaining books of the Pentateuch) is the era of the Exodus. With regard to a crucial date expressed in this era—"In the four hundred and eightieth year after the people of Israel came out of the land of Egypt, in the fourth year of Solomon's reign over Israel, in the month of Ziv, which is the second month, he began to build the house of the Lord" (I Kings 6:1)—there is again a discrepancy between the Hebrew text and the Septuagint. Other problems to be met with during this period are due to the obscurity of chronological data in the book of Judges and in I and II Samuel.

Dating by the kings of Israel

During the following period, the Bible uses the eras of the regnal years of monarchs (the kings of Judah, Israel, and Babylon) and of the Babylonian Exile. This period of biblical chronology likewise poses numerous problems, also the result of apparent inconsistencies of the synchronisms—*e.g.*, in the period from the accession of Rehoboam of Judah and of Jeroboam of Israel to the fall of Samaria "in the sixth year of Hezekiah [of Judah], which was the ninth year of Hoshea king of Israel" (II Kings 18:10) the reigns of the southern kingdom exceed those of the northern kingdom by 25 years.

The biblical data might be easier to harmonize if the occurrence of coregencies were assumed. Yet, as an ever-variable factor, these evidently would not lead to the determination of the true chronology of this period. Scholars therefore seek additional information from sources outside the Bible—*e.g.*, inscriptions on Assyrian monuments, which are dated by the so-called eponym lists. Substantial use also has been made of the data in the king list known as Ptolemy's Canon (compiled in the 2nd Christian century) commencing in 747 BC with the reigns of the Babylonian

kings (see above *Babylonian and Assyrian*). Scholars differ widely, however, in their interpretation of details, and numerous chronological problems remain unsolved. Only a few dates in this period can be fixed with any degree of confidence.

Reckoning by Persian kings

After the Babylonian Exile, as evidenced by the data in the Bible and the Aswān papyri, the Jews reckoned by the years of the Persian kings. The chronological problems of this period are caused by the apparent disorder in the sequence of events related in the biblical books of Ezra and Nehemiah and by the difficulty of identifying some of the Persian kings in question. For example, the King Artaxerxes of these books may stand for Artaxerxes I Longimanus (465–425 BC), for Artaxerxes II Mnemon (404–359/358 BC), or in the case of Ezra at any rate, for Artaxerxes III Ochus (359/358–338/337 BC).

From the Grecian period onward, Jews used the Seleucid era (especially in dating deeds; hence its name Minyan Shetarot, or "Era of Contracts"). In vogue in the East until the 16th century, this was the only popular Jewish era of antiquity to survive. The others soon became extinct. These included, among others, national eras dating (1) from the accession of the Hasmonean princes (*e.g.*, Simon the Hasmonean in 143/142 BC) and (2) from the anti-Roman risings ("era of the Redemption of Zion") in the years 66 and 131 of the Common (Christian) Era. Dates have also been reckoned from the destruction of the Second Temple (*le-ḥurban ha-Bayit*). The various styles of the latter, as also of the Seleucid era and of the era *anno mundi*, have often led to erroneous conversions of dates. The respective general styles of these eras correlate as follows: 3830 AM = year 381 of the Seleucid era = year 1 of the Era of the Destruction = year 69/70 of the Common (Christian) Era.

Earlier Jewish chronologies

The earliest Jewish chronologies have not survived. Of the work of the Alexandrian Jew Demetrius (3rd century BC), which deduced Jewish historical dates from the Scriptures, only a few fragments are extant. In the *Book of Jubilees*, events from the creation to the Exodus are dated in jubilee and sabbatical cycles of 49 and 7 years, respectively. Scholars differ as to the date and origin of this book. The era of the creation therein is unlikely to have been other than hypothetical.

The earliest and most important of all Jewish chronologies extant is the *Seder ʿolam rabbaʾ* ("Order of the World"), transmitted, according to Talmudic tradition, by Rabbi Yosi ben Halafta in the 2nd century AD. The author was possibly the first to use the rabbinic Era of the Creation. His chronology extends from the creation to Bar Kokhba in the days of the Roman emperor Hadrian (2nd century AD); but the period from Nehemiah to Bar Kokhba (*i.e.*, from Artaxerxes I or II to Hadrian) is compressed into one single chapter. The Persian phase shrinks to a mere 54 years. The smaller work *Seder ʿOlam zuṭa* completes the *Rabbaʾ*. It aims to show the Babylonian exilarchs as lineal descendants of David.

Megillat taʿanit ("Scroll of Fasting"), although recording only the days and months of the year without the dates of the years, is nevertheless an important source for Jewish chronology. It lists events on 35 days of the year that have been identified with events in five chronological periods: (1) pre-Hasmonean, (2) Hasmonean, (3) Roman (up to AD 65), (4) the war against Rome (65–66), and (5) miscellaneous. The authors, or rather the last revisers, are identified with Zealots guided by Hananiah ben Hezekiah ben Gurion and his son Eliezer. (E.J.Wi.)

Greek. As the cities of ancient Greece progressed to their classical maturity, the need arose among them for a chronological system on a universally understood basis. In the archaic period, genealogies of local monarchs or aristocrats sufficed for the historical tradition of a given area, and events were associated with the lifetimes of well-known ancestors or "heroes." The *synoikismos* (founding of the united city) of Athens took place "in the time of Theseus"; the Spartan ephorate (chief magistracy) was established "in the reign of King Theopompus." When the city-states adopted annual magistracies, the years were designated by the eponymous officials—"in the archonship of Glaucippus" or "when Pleistolas was ephor." This was the local usage throughout classical and Hellenistic Greece, the title of the magistrate varying in different cities. Sometimes tenure of a priesthood provided the chronological basis, as at Argos, where years were dated as the *n*th of the (named) priestess of Hera. The correctness of the series was a matter first of memory and later of careful record. The list of annual archons at Athens was known back to 683 BC (in modern terms). Lists of dynasties also amounted to recorded folk memory, and in all genealogical reckoning there is a point, for modern critics, at which acceptable tradition shades into myth. Corruption of the records was introduced through error or political design, and traditions often conflicted.

Dating by local records

Chronology became subject to systematization when cities felt a national need for accurate clarification of their past. In literature the growth of historiography initiated a search for a method of dating that could be universally applied and acknowledged. In the 5th and 4th centuries, local historians used local magistracies as their framework; research was devoted to rationalization of conflicting traditions and production of definitive lists. Charon of Lampsacus, perhaps in the early 5th century, compiled a record of Spartan magistrates; Hellanicus of Lesbos, author of the earliest history of Athens, wrote on the priestesses of Argos. Lists of victors in the great Olympic games were valid for all Greece, pointing the way to the widely accepted reckoning by Olympiads (see below). The Athenian Philochorus was the latest (early 3rd century BC) of compilers of Olympionikai.

The 5th-century historian Herodotus relied for his chronology principally upon the reckoning by generations used by his informants, conventionally accepted as showing three generations to a century. In some cases a 40-year, or other, reckoning was used, and varying traditions sometimes produced difficulty of synchronism. Thucydides, writing "contemporary" history, recognized the chronological problems involved. He dated the beginning of the Peloponnesian War by the Athenian, Spartan, and Argive systems and thenceforward marked the passage of time by seasonal indications. Synchronization was not helped by the fact that the official year began at different times in different cities. In later historical writing the impossibility of accurately coordinating the Athenian and Roman years resulted in serious chronological difficulties.

Dating by the Olympiads

The system of dating by Athenian archons came to be recognized outside Attica as of wider value, but, in the Hellenistic period, Alexandrian scholarship, represented especially by Eratosthenes of Cyrene, the "father of chronology," was instrumental in promoting the use of the Olympiads as an acceptable system, reckoning a four-year period from each celebration of the Olympic Games. Timaeus of Tauromenium (*c.* 356–260 BC) was the first historian to employ it, but it was little used outside historical writing. Aristotle had been concerned to identify the generation of the first Olympiad, accepted as 776 BC on modern reckoning. For convenience, the beginning of the Olympic year was equated with the summer solstice, when the Athenian year also began. This makes it generally necessary for a Greek year to receive a double date in modern terms (*e.g.*, the death of the philosopher Epicurus in 271/270 BC). Eratosthenes' system produced tables of dates, from which, for example, the fall of Troy could be dated to 1184/83 BC. The "Parian Marble" of 264/263 BC is an inscribed record of events from the time of Cecrops, first king of Athens, reckoning years between the date of the inscription, fixed by the Athenian archon, and each event concerned. Some cities inscribed lists of their eponymous magistrates; the Athenians were the first to do so *c.* 425 BC. A list from Sicilian Tauromenium originally spanned some 300 years. The regnal years of the Hellenistic monarchs or the count from a fixed event (a city foundation or refoundation) also provided acceptable chronological reckoning often useful for more than contemporary or local purposes.

The use of these chronological possibilities is best seen in historians using the annalistic method, of whom Diodorus Siculus is most notable. In the Christian period, Eusebius, followed by St. Jerome, began the work of reconciling all these indications to the Judaic tradition and produced

the foundation of chronology in terms of the Julian calendar upon which modern historians have constructed their framework.

For modern scholarship the problem, in E.J. Bickerman's words, is "how we know Caesar was assassinated on March 15, 44 BC." Before 480 BC, no date can be precise in terms of the Julian calendar unless confirmed by astronomical phenomena. Archaic chronology relies upon the typology of Corinthian pottery in relation to the foundation dates for Greek colonies in Sicily implicit in Thucydides, book VI. Julian dates given for this period (*e.g.*, for the tyranny of Peisistratus in Athens) stem from a complex combination of ancient chronographic tradition with modern archaeology, acceptable only with appropriate reserve. Literary tradition gives the succession of Athenian archons from 480 to 294 BC. The regnal, era, and Olympiad years also provide dates within a 12-month period. Closer dating is seldom possible unless the sources give precise information in calendric terms, as occasionally in literature and regularly in Athenian and Egyptian public documents. Even these are not translatable into Julian months and days unless coordinated with knowledge of contemporary solar or lunar phenomena and of possible official interference with the calendar.

(A.G.W.)

Roman. The establishment of a sound chronology for Roman history, as for Greek, depends on the assessment of the evidence available, which falls into two categories—literary and archaeological.

Literary evidence. Although by the late 3rd century BC the Greek mathematician Eratosthenes was working on the systematization of chronography and a series of learned historians had used the documentary method—*e.g.*, for Roman history, Timaeus of Tauromenium, to whom are probably due many of the synchronizations of Roman history with the Greek Olympiads—unfortunately this tradition of documentation and concern for chronology did not immediately pass over into Roman historiography. According to Cicero in *De oratore*, the earliest Roman historians did no more than "compile yearbooks"—for example, Fabius Pictor in the late 3rd century BC, Lucius Calpurnius Piso in the 2nd, and the so-called Sullan annalists in the 1st. Of these authors it is possible to judge only at second hand, and only those of the 1st century were much used directly by the historians whose work survives in any quantity, notably Livy, Dionysius of Halicarnassus, and Diodorus Siculus. In these authors, as in other 1st-century historians such as Sallust, there is little concept of documentation or research other than comparison of literary sources; for none was chronology a direct concern, and in many cases dramatic effectiveness took priority over fidelity to truth. Apart from the Greek Polybius, who treated the rise of Roman power in the Mediterranean from 264 to 146 BC, it was not until Cicero's time that the conception of historical scholarship developed in Rome. Cicero's friend Atticus not only was concerned to draw up a chronological table in his *Liber annalis* but had undertaken research to that end, and the great scholar Marcus Terentius Varro and a little later the learned Marcus Verrius Flaccus produced a vast body of erudite work, nearly all lost. To this source must probably be ascribed the *Fasti Capitolini*, a list of magistrates from the earliest republic to the contemporary period, set up near the *regia* (the office and archive of the *pontifices*, or high priests), perhaps on the adjacent Arch of Augustus, at the end of the 1st century BC. This work, since it is based on inscriptions, is sometimes given precedence over literary evidence, but, since it is a compilation, it is still subject to serious error.

Magistrate lists

Sources used by Roman historians. The traditionally early extant bodies of law, such as the Twelve Tables from the early republic, were of little chronological value, and juristic *commentarii* were liable to mislead through their zeal for precedent, while Cicero, in spite of Polybius' claim to have inspected early treaties preserved in the Capitol definitely states that there were no public records of early laws. A source frequently referred to is the *Annales maximi*, a collection made about 130 BC of the annual notices displayed on a white board by the *pontifices* and containing notes of food prices, eclipses, etc. Dionysius of Halicarnassus implied that they gave a date for the foundation of the city but was reluctant to accept their authority; and one of the eclipses is referred to by Cicero as being mentioned also by Ennius, but unfortunately the number of the year "from the foundation of the city" is corrupt in the text. Although it is possible to calculate the dates of eclipses astronomically in terms of the modern era, it is difficult to link these to Roman chronology because of the uncertainty of the figures and because of the confused state of the Roman calendar before the Julian reform (see CALENDAR: *Early calendar systems: The early Roman calendar*). Another difficulty is that the early records may have been burned in 390 BC when Celtic tribes sacked the city; also they would probably have been largely unintelligible if authentic.

Livy quoted the 1st-century annalist Gaius Licinius Macer as having found in the temple of Juno Moneta "linen rolls" giving lists of magistrates; but he also said that Macer and Quintus Aelius Tubero both cited the rolls for the consuls of 434 but gave different names. In any case, it is unlikely that the list could have been older than the temple, which dates from 344 BC. It is clear that the chief sources for the lists were the pedigrees of prominent Roman families, such as the Claudii Marcelli, Fabii, and Aemilii, drawn up by Atticus; but Cicero and Livy agree that tendentious falsifications had in many cases corrupted the records, and other suspicious facts are the appearance of obviously later or invented *cognomina*, or third names, and of plebeian gentile names for the earliest period, when only patricians bore them. Many scholars, however, accept the general authenticity of the lists—one reason being the appearance in them of extinct patrician families—but prefer Livy's version to that of the Capitoline lists, which show signs of late revision, often give names in incorrect order, and contain other anomalies.

The question, therefore, remains whether Roman chronography was dependent on the lists of magistrates or whether these were adapted to fit other known datings. The apparent advantages of the existence of a terminal date, the "foundation of the city," is illusory for Roman chronology, since it depended on back reckoning and was not agreed even in antiquity. Various ancient scholars each assumed a different date. Each computed his date by adding a different number of years of kingly rule from the foundation of the city to his estimate of the date of the foundation of the republic. This, in turn, was presumably computed by counting back over the yearly lists of magistrates. There may have been traditions about the intervals between certain events in early Roman history, but the frequently accepted reckoning of 244 years of kingly rule seems to be a calculation based only on the conventional 35-year generation for the rule of the seven legendary kings. Polybius claimed that the dating of the first republican consulship to 508/507 BC could be substantiated by an extant copy of a contemporary treaty. Combined with the traditional kingly period, this would give a foundation date of 751–750, reckoned inclusively, and 752–751, exclusively (Cato's date). The chronological scheme worked out by Varro added two years of nonconsular rule, thus the foundation of Rome was put in 754/753 and the beginning of the republic in 510/509. Varro's dates became standard for later Romans and are sometimes also used by modern scholars in a purely conventional sense. But it remains uncertain whether the dating depended on the magistrate lists or whether these were "doctored" to synchronize with given dates or intervals, whether these were traditional or calculated in some other way. Anomalies such as Livy's five-year anarchy 15 years after the Gallic invasion, Diodorus' repetition of magistrates' names, and the "dictator years" in the lists are perhaps attempts to synchronize the various pedigrees.

Fixing the terminal date

Contribution of archaeology. Archaeology can provide many dates useful to the detailed study of Roman history, especially from coins and inscriptions, but, for the general scheme of early chronology, its value is largely negative. It shows, for example, that Rome evolved over a lengthy period and was not really "founded," though a "foundation" date might perhaps refer to the first common celebration

of the Septimontium, or festival of the seven hills; again, if that dating is dependent on the seven kings, archaeology shows that the tradition about them, though it may preserve genuine names and events, is largely legendary.

Datings after the 1st century BC. In this better documented period, datings to consul years, or later to the years of tribunician power of the emperors, are normally intelligible, despite a few notorious cruxes, although up to the Julian reform the state of the calendar has always to be taken into account. In parts of the empire, however, different eras were used—*e.g.*, that of the Seleucids—and from the 4th century AD dates were often calculated in terms of the years of the indiction, a 15-year cycle connected with the levying of taxes, a method that continued in use for many centuries in spite of difficulties, such as lack of synchronization among the various provinces.

(Ed.)

The Christian Era

Christian. The Christian Era is the era now in general use throughout the world. Its epoch, or commencement, is January 1, 754 AUC (*ab urbe condita*—"from the foundation of the city [of Rome]"—or *anno urbis conditae*—"in the year of the foundation of the city"). Christ's birth was at first believed to have occurred on the December 25 immediately preceding. Years are reckoned as before or after the Nativity, those before being denoted BC (before Christ) and those after by AD (*anno Domini*, "in the year of the Lord"). Chronologers admit no year zero between 1 BC and AD 1. The precise date of commencing the annual cycle was widely disputed almost until modern times, December 25, January 1, March 25, and Easter day each being favoured in different parts of Europe at different periods.

The Christian Era was invented by Dionysius Exiguus (*c.* AD 500–after 525), a monk of Scythian birth resident in Italy; it was a by-product of the dispute that had long vexed the churches as to the correct method of calculating Easter. Many churches, including those in close contact with Rome, followed 95-year tables evolved by Theophilus, bishop of Alexandria, and by his successor, St. Cyril; but some Western churches followed other systems, notably a 532-year cycle prepared for Pope Hilarius (461–468) by Victorius of Aquitaine. In 525, at the request of Pope St. John I, Dionysius Exiguus prepared a modified Alexandrian computation based on Victorius' cycle. He discarded the Alexandrian era of Diocletian, reckoned from AD 284, on the ground that he "did not wish to perpetuate the name of the Great Persecutor, but rather to number the years from the Incarnation of Our Lord Jesus Christ."

Somehow Dionysius reckoned the birth of Christ to have occurred in 753 AUC; but the Gospels state that Christ was born under Herod the Great—*i.e.*, at the latest in 750 AUC. Dionysius' dating was questioned by the English saint Bede in the 8th century and rejected outright by the German monk Regino of Prüm at the end of the 9th. Nevertheless, it has continued in use to the present day, and, as a result, the Nativity is reckoned to have taken place before the start of the Christian Era.

The new chronology was not regarded as a major discovery by its author; Dionysius' own letters are all dated by the indiction (see below). The use of the Christian Era spread through the employment of his new Easter tables. In England the Christian Era was adopted with the tables at the Synod of Whitby in 664. But it was the use, above all by Bede, of the margins of the tables for preserving annalistic notices and the consequent juxtaposition of historical writing with calendrical computations that popularized the new era. Outside Italy it is first found in England (in a charter of 676) and shortly afterward in Spain and Gaul. It was not quickly adopted in royal diplomas and other solemn documents, however, and in the papal chancery it did not replace the indiction until the time of John XIII (965–972). The Christian Era did not become general in Europe until the 11th century; in most of Spain it was not adopted until the 14th and in the Greek world not until the 15th.

Of the alternative chronologies used by Christians, the most important were: (1) the indiction, (2) the Era of Spain, and (3) the Era of the Passion. The indiction was a cycle of 15 years originally based on the interval between imperial tax assessments but during the Middle Ages always reckoned from the accession of Constantine, in 312. Years were given according to their place in the cycle of 15, the number of the indiction itself being ignored. This chronology was the most widespread in the early Middle Ages, but its use diminished rapidly in the 13th century, although public notaries continued to use it until the 16th. The Era of Spain was based on an Easter cycle that began on January 1, 716 AUC (38 BC), marking the completion of the Roman conquest of Spain. First recorded in the 5th century, it was in general use in Visigothic Spain of the 6th and 7th centuries and, after the Arab invasions, in the unconquered Christian kingdoms in the north of the Iberian Peninsula. It was abolished, in favour of the Era of the Incarnation, in Catalonia in 1180, in Aragon in 1350, in Castile in 1383, and in Portugal in 1422. The Era of the Passion, commencing 33 years after that of the Incarnation, enjoyed a short vogue, mainly in 11th-century France. (Ed.)

Muslim. Unlike earlier chronological systems in use before Islām, Islāmic chronology was instituted so soon after the event that was to be the beginning of the Muslim era that no serious problems were encountered in its application. According to the most reliable authorities, it was 'Umar I, the second caliph (reigned 634–644), who introduced the era used by the Muslim world. When his attention was drawn by Abū Mūsā al-Ash'arī to the fact that his letters were not dated, 'Umar consulted with men at Medina and then ordered that the year of the *hijrah* (hegira), the Prophet's flight from Mecca to Medina, be taken as the beginning of an era for the Muslim state and community. According to the Muslim calendar, the *hijrah* took place on 8 Rabī' I, which corresponds to September 20, 622 (AD), in the Julian calendar. But, as Muḥarram had been already accepted as the first month of the lunar year, 'Umar ordered that (Friday) 1 Muḥarram (July 16, 622) be the beginning of the reckoning. It is generally accepted that this was done in AH 17 (*anno Hegirae*, "in the year of the Hegira").

The era of the *hijrah*

There are a few points in connection with this that deserve mention: first, there is no real agreement on the exact date of the *hijrah*—other dates given include 2 and 12 Rabī' I; second, the year in which 'Umar issued the order is a point of contention—the years 16 and 17 are sometimes given; third, some people have ascribed the use of the chronology to the Prophet himself. According to some sources, the *hijrah* date was first used by Ya'lā ibn Umayyah, Abū Bakr's governor in Yemen. This sounds somewhat plausible because Yemenis were probably used to affixing dates to their documents. There is, however, a consensus among workers in the field that 8 Rabī' I was the day of the *hijrah*, that 'Umar instituted the use of the date for the new era, and that this was done in AH 17. The choice of the *hijrah* as the beginning of the epoch has two reasons. On the one hand, its date had been fixed; on the other, 'Umar and his advisers must have recognized the importance of the migration—Islām had become, as a result, a religion and a state.

Before the introduction of the new epoch, the Arabs had been acquainted with chronologies used by their neighbours, the Seleucids and the Persians. In Yemen the practice of dating had been perfected to the extent that inscriptions show the day, the month, and the year. In Mecca the "year of the Elephant," supposedly coinciding with the birth of the Prophet, had been in use. For the period between the migration and the institution of the new epoch, the Muslims of Medina resorted to naming the year after local events—"the year of the order of fighting" and "the year of the earthquake," etc.

The lunar year was adopted by the Muslims for the new chronology. In this there was hardly any innovation insofar as Arabia was concerned.

The chronology introduced by 'Umar was adopted throughout the Muslim world, although earlier epochs continued in use in outlying provinces. Muslim historians, annalists, and chroniclers met with difficulties when writing their books on pre-Islāmic history. No practice had as yet developed for pre-*hijrah* dating; therefore, when writing about the history of various lands in pre-Islāmic times,

authors resorted to the use of chronologies previously in existence there (*e.g.*, Persian, Indian, Seleucid, Alexandrian). For the histories of the area under Islām, writers used only Muslim chronology, while non-Muslim authors (*e.g.*, Bar Hebraeus) used the Seleucid and the *hijrah* dates when discussing events pertaining to provinces that had been Byzantine and therefore still had fairly large groups of Christians.

The era of the *hijrah* is in official use in Saudi Arabia, the two Yemens, and in the Persian Gulf area. Egypt, Syria, Jordan, Morocco, Algeria, Libya, and Tunisia use both the Muslim and the Christian eras. Many Muslim countries, such as Turkey, Nigeria, and Pakistan, use the Christian Era.

Variants of the *hijrah* era

Within the general uniformity of applying the *hijrah* era proper, there existed differences, some of which were the result of earlier pre-Islāmic practices; others were the result of continuous contacts of Muslim countries with their European neighbours, with whom they had economic as well as political relations. An example of the former was the work of the 'Abbāsid caliph al-Mu'taḍid, who brought the Nowrūz (Persian New Year's Day) back to date in keeping with the agricultural activities of the community. Maḥmūd Ghāzān introduced the Khānian era in Persia in AH 701, which was a reversion to the regnal chronologies of antiquity. It continued in use for some generations, then the ordinary *hijrah* era was reintroduced. A similar step was taken by Akbar when he established the Ilāhī era, which began on Rabī' II 963 (February 13, 1556), the date of his accession; the years were solar.

Two Muslim countries, Turkey and Iran, introduced more drastic changes into their chronology because of European influences.

In Turkey the Julian calendar was adopted in AH 1088 (AD 1676–77) and used solar months with *hijrah* dating. The year was officially called the Ottoman fiscal year but was popularly known as the *marti* year, after *mart* (Turkish for March), which was the beginning of the year. Under Mustafa Kemal Atatürk, the Gregorian calendar and the Christian Era were officially adopted in Turkey (1929). Iran also adopted a solar year; the names of the months in its calendar are Persian, and the era is still that of the *hijrah*. (N.A.Z.)

Pre-Columbian American. *Maya and Mexican.* The lowland Maya had a 365-day year formed of 18 "months." Each month consisted of 20 days, plus five "nameless" days, which the Maya considered an extremely dangerous and unlucky period and during which activities were kept to a minimum. Leap days were not intercalated.

Reckoning was not by those years but by *tuns* (360 days) and their multiples of 20: *katuns* (20 *tuns*), *baktuns* (400 *tuns*), *pictuns* (8,000 *tuns*), *calabtuns* (160,000 *tuns*), and *kinchiltuns* (3,200,000). In practice, the last three were seldom used. The *tun* comprised 18 *uinals*, each of 20 *kins* (days), but these did not coincide with the equivalent divisions of the 365-day year. The Maya normally carved or wrote these in descending order; students transcribe them in Arabic numerals—*e.g.*, 9.10.6.5.9 represents nine *baktuns*, ten *katuns*, six *tuns*, five *uinals*, nine *kins*.

With this system, current dates were related to the start of the Maya era, which, because of the Maya system of re-entering cycles, marked both the end of 13 *baktuns* (written 13.0.0.0.0) and the start of another cycle of *baktuns* and perhaps commemorated a re-creation of the world, the *baktun* about to enter being numbered 1, not 14. Because of the construction of the calendar, this start of the era happened to be day 4 Ahau falling on the eighth day of the month Cumku.

Such reckonings are called Initial Series, or Long Counts, the former because they usually stand at the start of an inscription (see CALENDAR). For example, the combination day 8 Muluc, falling on second of Zip (third month), recurs every 52 years, but the Initial Series (here 9.10.6.5.9 8 Muluc 2 Zip) pinpoints its position. The next occurrence, 52 years later, would be 9.12.19.0.9 8 Muluc 2 Zip. Each unit had its own glyph (or symbolic character), with appropriate number (normally a dot for 1 and bar for 5) attached.

A shorter dating system was by "Period Endings"—that is, by recording the ending of the current *baktun, katun,* or *tun.* Thus, day 13 Ahau and month position 13 Muan with 13 *tuns* added is an abbreviation of 9.17.13.0.0 13 Ahau 13 Muan, a combination that will not repeat for over 900 years (949 *tuns*). A still shorter but less precise method was to give the day and its number ending the current *katun.*

Several Maya dates were commonly linked to Initial Series or Period Endings by series of additions or subtractions—a glyph signifying count indicated forward or backward by secondary attachments.

Dates were normally reckoned from the 4 Ahau 8 Cumku base, nearly 4,000 years before most inscriptions, but some calculations ranged far into the past and a few into the distant future. One reaches backward nearly 1,250,000 years, but the deepest probings of eternity are embodied in texts that seemingly record positions respectively 90,000,000 and 400,000,000 years ago. Although the interpretation of these last computations is disputable, the Maya certainly thought in millions of years a millennium before Europe discarded the view that the world was only some 6,000 years old.

Maya conception of time

The Maya conceived of time as a journey through eternity in which each deified number—all time periods and their numbers were gods—carried his period on his back supported by a tump line. Each evening the procession rested. Next morning, carriers whose period was completed were replaced. For instance, if the *uinal* and *kin* numbers were 15 and 19 respectively, the new carriers would be the deified 16 and 0 (the latter because *kin* numbers go no higher than 19). Other period numbers would journey on until it came time to change the *tun* carrier. Much ritual and imagery grew out of this concept of the march of time; sculpture illustrates bearers lowering their burdens at journey's end.

Correlation of the Maya calendar with ours depends on several factors. First, the 260-day almanac still functions in some Maya villages in the Guatemalan highlands. As there is excellent evidence it has neither gained nor lost a day since the Spanish conquest, despite strong Spanish efforts to suppress it, one may reasonably assume no break under the more favourable pre-Columbian conditions. Lunar and other data support such a view. Second, month positions in Yucatán and southern Petén at the Spanish conquest also are reliably correlated to the day with the present Western calendar. Third, the combined day and month parts of the Maya calendar are in day-for-day agreement with the present Western calendar within a 52-year span (after that given day and month positions repeat). The *katun* (specifically, 13 Ahau) current at the Spanish conquest is, however, known, thereby fixing any day and month position in a longer range of 260 years because a named *katun* repeats only after 260 *tuns.* Those conditions produce a correlation of the two calendars that is either correct to the day or is 260 or even 520 years wrong, since historical evidence does not specify which particular *katun* 13 Ahau coincided with the Spaniards' arrival. Fourth, such factors as astronomy (Maya records of heliacal risings of Venus and of many dates with moon age stated), pottery sequences, architectural changes (less reliable), and data from neighbouring areas govern choice of the applicable *katun* 13 Ahau. Weight of evidence led to wide acceptance of the Goodman–Martínez–Thompson correlation that equates 13.0.0.0.0 4 Ahau 8 Cumku, start of the Maya era, with August 10, 3114 BC, and the Classic period with AD 300 to 900. Fifth, when the carbon-14 dating technique was first applied to the problem, various difficulties attendant on the use of new techniques and failure to take into account that a tree dies year by year from its centre outward (so that a sample from the core might give a date well over a century before felling) distorted readings, producing results favourable to the correlation making Maya dates 260 years earlier. Now, with better technique and averaging of many "runs" of samples of latest growth from beams at Tikal with secure Maya dates, carbon-14 readings overwhelmingly support the Goodman–Martínez–Thompson correlation.

The only other Middle American calendar with a known era is that of the Cakchiquel of highland Guatemala.

The system was vigesimal: *kih*, day; *uinak*, 20 days; *a*, 400 days; and *may*, 8,000 days. The 400-day "year" ran concurrently with the 260-day almanac, which, in turn, synchronized with all other Maya almanacs. Like the 360-day *tun* of the lowlands, the 400-day *a* was the counting unit, for reckoning was always in multiples of the *a*, never by days, as in our Julian calendar. *May* signifies twenty, and is so named because it comprised 20*a*. At the arrival of the Spaniards, reckoning was from a revolt in AD 1493. Earlier eras may be postulated, but inscribed calendrical texts are lacking in Cakchiquel territory.

Aztec. The Aztec and related peoples of central Mexico employed the cycle of 52 years, constructed, like its Maya equivalent, of concurrent 365-day years and 260-day cycles, any position of the former coinciding with a given position of the latter only at 52-year intervals. Again leap days were not used. At completion of the 52 years, known as "binding of the years," elaborate ceremonies were held to avert destruction of the world expected on that occasion. The last occurrence before the Spanish conquest was in AD 1507. Although the last creation of the world was designated by a day name, neither that nor any other was in general use in central Mexico as the start of an era. Aztec reckoning is normally from their arrival in the Valley of Mexico, supposedly the year 1 Flint (AD 1168).

There is much confusion in placing events in Mexican history because no system of distinguishing one 52-year cycle from another was employed except by writing every year glyph throughout the period covered, a clumsy arrangement. Each year was named for either its last day (omitting the five-day unlucky period) or for the last day of the fifth month (both choices have distinguished supporters). In either case, only four day names (House, Rabbit, Cane, and Flint), each with its accompanying numeral, could designate a year. The Spanish conqueror Hernán Cortés seized the Aztec capital in 1521, year 3 House, but some past event, also assigned to a year 3 House but unlocated in a full sequence of years, might refer to AD 1261, 1313, or 1365, etc. Month positions were rarely given in chronological statements.

Peoples of Oaxaca and the Isthmus of Tehuantepec. Pictorial books of the Mixtec of Oaxaca record events in the lives of ruling families covering seven centuries, but, again, happenings are fixed only by the day on which each occurred and the year in which the day fell. Sequence is usually clear, but at times there is doubt as to which 52-year period is meant when parenthetical material, such as life histories of secondary characters, is inserted.

No era is recognizable. A clouded entry concerning the descent to Earth of the Sun and Venus, perhaps assignable to AD 794, is a logical starting point, but other entries are earlier.

Little is known of the calendar of the Zapotec, neighbours of the Mixtec. Years began on a different set of days, and glyphs differ from those of Mixtec and Aztec. Months are not recorded on monuments, which are numerous, and no chronological system has survived. Most Zapotec texts are early.

Rare inscriptions in western Chiapas, southern Veracruz, and the Guatemalan Pacific coast resemble the abbreviated lowland Maya Initial Series used in script and on a single sculpture in that numerical bars and dots are in a vertical column with period glyphs and month signs suppressed, clearly place numeration, that is, the value of each unit was shown by its position in the column. The linguistic affiliation of their sculptors is unknown.

All texts are either fragmentary or damaged; the two complete ones, unlike Maya Initial Series, open with days signs (and different ones at that). If, as one may reasonably assume, the series of bars and dots departed from those day signs, a fixed era is questionable. Nevertheless, some scholars postulate use of the Maya era (13.0.0.0.0 4 Ahau 8 Cumku). This little understood system may have been ancestral to the Maya Initial Series, the Maya perhaps developing a fixed era, for they alone seem to have been interested in an exact chronological system.

(E.Th.)

DIPLOMATICS

Diplomatics, broadly speaking, is the study of documents. The term is derived from the Greek word *diploma*, meaning "doubled" or "folded." Besides the documents of legal and administrative import with which it is properly concerned, diplomatics also includes the study of other records such as bills, reports, cartularies, registers, and rolls. Diplomatics is therefore a basic and not simply an auxiliary historical science. This article deals with its development and practice in the Roman Empire and in Europe. During Roman antiquity certain documents containing different sorts of authorizations were engraved on a bronze diptych and then folded and sealed, in order to keep the contents secret—hence the term diploma. Rarely found during the Middle Ages, the word was used by the Renaissance Humanists to denote formal documents of ancient rulers. The interest in and description of such documents came to be called *res diplomatica* after the famous 17th-century work *De Re Diplomatica Libri VI*, by Jean Mabillon, a member of the scholarly Benedictine congregation of Saint-Maur. Mabillon's work first made the study of old documents a reputable science. The major task of diplomatics is to distinguish between genuine and false documents, and this involves detailed examination of their external and internal features. Diplomatic studies have been applied mainly to Western documents, usually medieval ones, because it requires less specialist training to analyze more recent documents.

History of the study of documents. *Medieval and Renaissance work.* The forging of documents took place on a vast scale during the earlier Middle Ages, partly because wars and disturbances so frequently upset possession and also because the increasing use of written records made it necessary for those whose title was, in fact, perfectly good in old unwritten "customary" law to give it written substantiation. Thus forgeries, partly intentionally honest, partly dishonest, occurred frequently, despite the fact that the Germanic tribes that settled in western Europe inherited, with other aspects of Roman law, the concept of forgery as a felony, which was soon also reinforced by the church's canon law. This legal concept of forgery was, however, mainly applied to cases concerning property or inheritance; and literary forgeries, such as the famous Donation of Constantine, which purported to be the gift by the Roman emperor Constantine I the Great (died 337) to Pope Sylvester I of spiritual primacy throughout the church and of temporal power in Italy, were not concerned. Serious critical efforts to detect forgery did not begin in the Middle Ages, although obvious forgeries might be challenged in the course of a dispute. As early as the 6th century, the Merovingian king Childebert II declared a charter recording the gift of land from himself to the Bishop of Reims a forgery on the simple ground that the royal official denied the signature on it to be his. Pope Innocent III (1198–1216) tried to establish infallible criteria for the detection of fraudulent papal documents, but knowledge of earlier documentary forms was totally inadequate. In the Renaissance the Humanists began to use philological and technical criteria; on these grounds Lorenzo Valla authoritatively pronounced the Donation of Constantine to be a forgery, though authenticity had already been questioned.

The forging of documents

Post-Renaissance scholarship. Three events in the 17th century forced the development of more sophisticated standards of evaluation. The Thirty Years' War in Germany led to endless legal conflicts, and in France the nobility engaged in a concerted action known as the *bella diplomatica* ("diplomatic wars") to assert their ancient privileges against royal absolutism. The decisive impetus, however, came from a much more particularist dispute. Daniel van Papenbroeck, a member of the Jesuit commission known as the Bollandists (from another member, Jean Bolland), which was charged with the publication of the *Acta Sanctorum* ("Acts of the Saints,"), finding that some monastic documents he inspected were forgeries, assumed (1675) that this was true of almost all early-medieval documents. Since most of the monasteries with which the documents had been concerned were of the Benedictine Order, the Benedictines resented the sugges-

tion, and Mabillon undertook to refute it. In his *De Re Diplomatica* (1681), Mabillon set out the fundamental principles of the science of verifying documents; Papenbroeck soon afterward acknowledged the correctness of his tenets. Nearly a century later, René-Prosper Tassin and Charles-François Toustain published their six-volume *Nouveau traité de diplomatique* (1750–65; "New Treatise on Diplomatic"), a work that surpassed Mabillon's only in its greater wealth of material. Another important event in the history of the science of diplomatics was the founding of the École des Chartes (an institute for the training of French archivists) in Paris in 1821. During the next decades important collections of early-medieval French documents were printed in the *Recueil des actes* by a variety of eminent editors. But the greatest advances were made by German and Austrian scholars, among whom Julius von Ficker investigated the differentiation between *actum* and *datum* (that is, between verbal legal procedure and its formal documentation), and Theodor von Sickel defined a basic technique of studying and comparing the script of charters and thus of identifying the individual notaries or scribes. The diplomas of the Carolingian and the German kings and emperors were edited in the series of the *Monumenta Germaniae Historica*, by members of the Institut für österreichische Geschichtsforschung (Institute of Austrian History Research), established by Sickel in 1854. Meanwhile, the *Regesta*, comprising short, synoptical condensations of the contents of papal documents down to 1198, published by Philipp Jaffé in 1851, gave a decisive momentum to the study of the papal chancery, while August Potthast covered the period from 1198 to 1304. Prominent scholars in the research of papal records in Germany at the beginning of the 20th century were Michael Tangl, Rudolf von Heckel, and, particularly, Paul Fridolin Kehr. In comparison with the amount of work done in France and Germany, historical scholarship in England long paid relatively little attention to legal, as opposed to literary, records. Although John Mitchell Kemble published his collection of Anglo-Saxon documents, the *Codex Diplomaticus Aevi Saxonici* (1839–48), an extensive study of Anglo-Saxon and Norman legal and administrative documents was delayed until the 20th century. Since then notable contributions have been made by scholars such as Helen Cam, H.W.C. Davis, Vivian Hunter Galbraith, Frank M. Stenton, Dorothy Whitelock, David Charles Douglas, and many others. Christopher Robert Cheney has made important contributions to the research of papal documents. In Italy Luigi Schiaparelli made vital contributions to the study of Lombard documents. From the 19th century, some study of documents has formed part of the medieval-history curriculum in most European universities.

Lack of study in England

Diplomatic method. *Types of documents.* Documents that have been preserved are either originals, drafts, or copies. Originals, of which many survive, are formal documents drawn up on the order of the sender or donor, and they were designated to serve the recipient or beneficiary as evidence of the transaction recorded. Handwritten copies of documents, made either before or after the deed was actually executed (sealed), are not classified as originals. If made before an "original," they were in fact rough drafts of it; if made afterward, they were copies. The particularly Anglo-Saxon method of chirography, however, gave the possibility of producing several "originals." By this process two or more specimens of a document were written on the same page of the vellum sheet, and the free space between the texts was filled in with the word *chyrographum* ("handwriting") or other words and symbols. Then the sheet was cut irregularly right through these words or symbols; the originals thus separated could later be reassembled, an exact fit being complete proof of authenticity. But to provide documents having the force of "originals," copies of the original were usually made and formally certified as such, by public notaries, or by high ecclesiastical or secular dignitaries. Copies certified in this way were accorded the same legal value as the originals. In practice, lack of critical judgment on the part of the certifiers often led to the certification of forged records. In documents known as transumpts, which recited earlier documents or charters as part of their text, it often happened that the earlier document was forged, but, being included in the new, it received validation. The original documents and copies considered above were issued at the request of the recipient or beneficiary or of his legal heir. It also happened quite often that the sender or donor wished for various reasons to retain a record of the documents issued by him. The chanceries (record offices) of secular rulers or great ecclesiastics therefore kept copies of outgoing documents in registers, and often of incoming documents, too. The popes were among the first to adopt the old Roman practice of keeping registers; although nearly all the earlier ones have been lost, an almost uninterrupted series of papal registers is extant from the pontificate of Innocent III onward. An important group of registers are the rolls kept by the medieval kings of England; the earliest extant rolls date from the 12th century. The keeping of registers in the chanceries of the French kings began about the year 1200, in Aragon about 1215, in Sicily under the Hohenstaufen emperor Frederick II (died 1250), and in the German imperial chancery from the early 14th century. Another manner of studying documents is in the formula books of the various chanceries. Notaries drawing up the various forms of medieval documents did not usually compose each new text afresh but, rather, copied from books in which such text formulas had been collected, a practice that can be traced back to Roman procedure. These model texts frequently contained only the legally relevant passages, while the individually applicable parts, such as names, figures, and dates, were either abridged or totally omitted. During the time of the Frankish kings, important collections were made, such as the *Formulae Marculfi* (early 8th century) and the *Formulae imperiales* (828–832). Significant collections of formulas serving as models for papal documents have been preserved from the 13th century.

Use of formulas

Classification of documents. The documents of the Middle Ages are usually classified under two groups: public documents, which are those of emperors, kings, and popes, and private documents, which comprise all others. Another way of classifying documents is according to whether they are evidentiary or dispositive. The former merely record a valid legal act already executed orally, while the actual issuing of the latter forms in itself the legal act. This distinction, found among Roman documents from the 3rd century AD onward, gradually ceased to exist after the early Middle Ages. After the collapse of the Carolingian empire in the 9th century, private documents lost much of their function and were replaced by simple memorandums about legal acts and the witnesses to them. It was not until the late 11th and early 12th centuries that sealed charters of high secular or ecclesiastical dignitaries were again gradually considered as dispositive. Papal documents can be classified mainly as either letters or privileges, and royal documents can be classified as diplomas or mandates. Privileges and diplomas give evidence of legal transactions designed to be of long duration or even of permanent effect, while mandates and many papal letters contain commands.

Physical appearance of documents. Documents were written on a variety of material. In antiquity there were documents of stone, metal, wax, papyrus, and, occasionally, of parchment, but only papyrus and parchment (and, very occasionally, wax) were used during the Middle Ages. From the 12th to the 13th centuries, paper also was sometimes available. Papyrus, made from the stem of the papyrus plant, was produced mainly in Egypt; after the Arab conquest of Egypt in the 7th century, the import of papyrus to Europe became difficult. The Merovingian kings wrote their documents on papyrus until the second half of the 7th century, and the popes did so until far into the 11th century. North of the Alps papyrus had finally disappeared by the 8th century, when it was replaced by parchment. Parchment was made from animal hides and was thus easier to obtain. In southern Europe it was made mainly from sheep and goat hides; the insides of the skin were thoroughly smoothed and calcined, while the hairy sides were left rougher. In central and northern Europe, parchment was usually made from calves' skins,

Introduction of parchment

and both sides of the hides were thoroughly smoothed and calcined. Paper came originally from China. During the 8th century AD, it spread to the Arab world and from thence to Byzantium, where it was manufactured from linen and was used from the 11th to the 13th centuries for imperial documents. After that time ordinary paper was used in the Byzantine Empire. In the West the use of paper, most common at first in southern Italy and Spain, had begun to spread by the beginning of the 12th century. Germany and southern France began to import paper from Spain and Italy in the 13th century, and soon afterward it had reached England by way of Bordeaux. But paper did not altogether replace parchment, which long remained in use, especially for solemn documents. The medium for writing was ink, generally a mixture of oak gall and copper vitriol. Originally black, ink made north of the Alps sometimes shows a reddish-brown hue, while that made in Italy may contain tinges of brown and yellow. Over the centuries most of these colours have lightened as a result of atmospheric conditions. The Byzantine emperors used purple ink for their signatures. This custom was occasionally taken over by the Lombard rulers of Italy and, later, by the Norman kings of Sicily. Another custom of Byzantine origin is the use of gold lettering.

Throughout the entire Roman Empire, the language used in documents was primarily Latin. Greek was also used, and, during the latter part of the 6th century AD, it slowly superseded Latin in the East. From then onward, Greek was the language of Byzantine documents until the end of the Byzantine Empire (1453). In the West, the collapse of the empire and the establishment of barbarian kingdoms led to a vulgarization of Latin, written as well as spoken.

Use of vernacular

Latin has always been used for papal documents and for most public and private charters, and it was used for international documents well into post-Renaissance times, until it was superseded by French as the language of diplomacy. In public and private documents, use of the vernacular alongside Latin gradually developed. Apart from its early and unique appearance in the documents of the Anglo-Saxons in England, no vernacular was used in charters before the 12th century. At the Norman Conquest (1066), use of Anglo-Saxon in English documents soon stopped, and no more vernacular was used there until some Norman French was introduced in the 13th century, and Middle English in the 15th century. There was an increasing use of the vernacular in Italian and French documents from the 12th century and in Germany from the 13th; but in medieval times Latin was never outstripped by the vernacular.

A correct assessment of the hand in which it was written is vital to ascertaining the provenance and authenticity of a document. Thus, the knowledge of paleography, different styles of ancient writing, is a skill essential to diplomatics. The broad basis of such knowledge begins with acquaintance with the general styles of writing current at particular times and places. This varied with the way the pen was held; whether the writing was cursive or had the letters formed separately; whether it was majuscule, all the letters being contained between a single pair of horizontal lines, or minuscule, with parts of the letters extending above and below the lines. There is a further distinction between what is called book hand and the business, or court, hand at one time used for documents.

In Europe the Roman capital letters, distinguished as rustic or square, uncial, and Roman majuscule and minuscule cursive, influenced all subsequent writing in the West. The Roman curial style (from the Curia, or papal court), used in the papal chancery until the 12th century, was a derivation of late Roman minuscule cursive. After the disintegration of the Western Empire, the Merovingian Franks used a Roman provincial script for their documents. Distinctive forms developed elsewhere, in Visigothic Spain and in Ireland. The Irish script, a half uncial (uncials are rounded letters) and a minuscule script, spread to Anglo-Saxon England and thence to the European continent. Under the Carolingian rulers, a particularly clear and attractive minuscule book hand (Caroline minuscule) was developed; modifications of this gradually became used in documents and eventually spread also to Italy, England, and Spain. A "Gothic," more pointed form of script developed since the 11th century in northern France and soon spread all over Europe, so that writing became more spidery in appearance. In the early years of the Renaissance, Italian scholars such as Poggio (Poggio Bracciolini) and Niccolò Niccoli developed a minuscule based on the Carolingian, and variants of this style were used by the Venetian Aldus Manutius and other pioneers of printing.

Abbreviations were used in both documents and books. Again, their particular characteristics would contribute to a correct assessment of the probable date and provenance of a document. Roughly two types were used: the suspension, involving the writing of only the first letter or syllable of a word; and the contraction, used first for Hebrew and Christian sacred names, the writing of only the first and last letter or letters of a word or syllable. The Carolingians sometimes used Tironian notes, a form of shorthand devised by Tiro, a freedman of the Roman orator Cicero.

From Roman times the two most important methods of validating documents were by appending the signature or the seal of the sender or promulgator. The practice of using seals for this purpose (and not merely to close a document) was carried over from imperial usage and, by the 8th century, was current among the Lombards and other Germanic tribes in western Europe. Until about the 8th century, the signature of the Merovingian ruler or his delegate was also required for the validation of public documents, but thereafter the seal alone, together with the recognition by a high chancery official, was held sufficient, the king's signature dwindling into a monogram or mere stroke (the "stroke of execution"). This change was probably accelerated because many medieval kings could not write. Thus, in England, King John sealed, and did not sign, the Magna Carta.

Seals

Seals were made of wax or of metal; if the latter, they were called bulls (hence, the use of this term for a certain group of papal documents). The Byzantine emperors used gold seals for their documents; Byzantine officials and ecclesiastics used lead and silver for their bulls. Papal seals were of lead or gold. Wax seals were increasingly used from about the 11th and 12th centuries, and wax was also used for the impression when, later, less formal documents were validated by use of the signet or privy seal. Seals could be two-sided, suspended from the document, or impressed upon it.

Form and content of documents. Normally, each document was divided into three distinct parts: the introduction (protocol), the main text (context), and the concluding formulas or final protocol. There were various subdivisions, and not all the parts here mentioned are necessarily found in every document. The introduction comprises, first, the invocation (*invocatio*) of God, either by name or through a symbol such as the cross; second, the superscription (*intitulatio*), giving the name and title of the sender; and third, the address (*inscriptio*), naming those to whom the document is directed, usually followed by a formula of greeting (*salutatio*). The actual text of the document can be divided into a number of parts. The first, known as the *arenga*, expresses in general terms the motive for the issue of the document. The notification (*promulgatio*), briefly explaining the legal purpose of the document, is followed by the *narratio*, or exposition of the particular circumstances involved. In the *dispositio* the donor or promulgator firmly declares his purpose ("I hereby decree" or "I hereby give"); this clause is the vital core of the document, its legal decree of enactment. There usually followed the *sanctio*, a threat of punishment should the enactment be violated. The main text concluded with the *corroboratio*, a statement of the means to be used for validation of the document. The final protocol consisted of subscriptions or lists of names of all those, such as the scribe, who took part in the issue of the document and of witnesses to the enactment. The date and place of issue are given, and the final sentence, the *apprecatio*, is a short prayer for the realization of the contents of the charter. At the bottom of the document, the signs of validation (the recognition, monogram, seal) were then added.

The date given on a document might be either that of legal enactment (*actum*) or that of the issue of the doc-

Problems of chronology

ument recording the (already performed) legal enactment (*datum*). The form in which dates are given in a document is of particular import in determining its provenance and authenticity. A wide variety of practices were followed at different places and times. For instance, days of the month could be given according to the old Roman system of calends, ides, and nones; by continuous counting throughout the month; or by reference to a saint's day. Years might be computed from the presumed time of the creation of the world; by the Roman indiction, a 15-year cycle; by the names of officiating Roman consuls; by regnal years of emperor, king, or pope; or from the birth of Christ. Moreover, there were also a variety of ways to determine when the year began.

Development and characteristics of chanceries. *The Roman and Byzantine empire.* Rulers, all of whom needed to issue directives and edicts, developed writing offices, or chanceries, in which formal documents were drawn up. The Roman imperial chancery, called the Office of Letters (*ab epistulis*), was subdivided into a Greek and a Latin department. In the 5th century four letter offices existed, all under the ultimate control of the *magister officiorum* ("master of offices"): the *scrinium epistolarum* ("letter office") handled mainly foreign, legal, and administrative affairs; the *scrinium libellorum* ("petitions office") handled petitions and investigations; the *scrinium memoriae* ("memorandum office") composed shorter imperial decrees; and the *scrinium dispositionum* dealt with administration. From the 4th century, a group (*schola*) of notaries had come into being, some of whom served the emperor as personal secretaries. Two centuries later a special confidential (*a secretis*) secretary existed. In the Byzantine Empire in the 8th to 9th centuries, the *scrinium epistolarum* and the *scrinium libellorum* merged to form a new department under the *koiaistor* (a high palace official), while the secretaries had all come under the office of the *protoasekretis* (head of the secretaries). An official called the *mystikos* handled the emperor's secret correspondence. In preparing edicts or other laws, the *koiaistor,* after consulting the emperor, made a first draft of the bill, had the official copy drawn up by notaries, and then verified its accuracy before it was validated. From the 9th century onward other high court officials participated in the validation of Byzantine charters.

The important governmental documents of the late Roman and early Byzantine empires include laws, edicts, decrees (imperial decisions concerning civil and penal law), and rescripts (the emperor's replies to inquiries from corporate and administrative bodies or private persons). In the Byzantine era documents concerning more day-to-day affairs can be grouped under the headings of foreign letters, privileges, and administration. Foreign letters include correspondence with other rulers, treaties (regarded not as an agreement between equals but as an act of grace or privilege granted by the emperor, and made out as such), and letters accrediting imperial ambassadors. The most solemn and splendid form of privilege was the *chrysobullos logos,* so named because the word *logos,* meaning the emperor's solemn word, appeared in it three times, picked out in red ink. Written in the carefully embellished chancery script reserved for the emperor's personal documents, the text consists of the usual parts—that is, the *invocatio, intitulatio, inscriptio, arenga, narratio, dispositio, sanctio,* date, and the *subscriptio.* It was sealed with the golden bull.

Imperial privileges

From about the 12th to the mid-14th century, a simplified form, the *chrysobullon sigillion,* was used for privileges of lesser importance. It was not signed by the emperor himself but was held to be validated by the insertion, by the emperor, in red ink of the *menologema,* a statement of month and indiction. It, too, was sealed with a golden bull. The administrative documents of the Byzantine imperial chancery include the *prostagma,* or *horismos,* a plain and short document known since the beginning of the 13th century. If directed to a single person, the document starts out with a short address, but, in all other cases, it begins immediately with the *narratio,* followed by the *dispositio.* The emperor replaced his signature with the *menologema.* Unlike the privileges, this document was not rolled up but, instead, was folded, and then closed by means of a wax seal stamped with the imprint of the imperial signet ring.

In addition to those emanating from the imperial offices, there were other types of documents issued in the Byzantine Empire. These include those issued by despots and imperial officials and, in the ecclesiastical sphere, by patriarchs and bishops. There were also private documents. The documents issued by the despots carried a silver seal, showing their intermediate status between that of imperial documents sealed with the gold seal and that of documents drawn up by imperial officials and sealed with lead bulls. Documents issued by imperial officials were simpler. They lacked the protocol, and the personal signature of the issuing official was written in black ink. The detailed date comprises the *menologema.* The documents of the Byzantine patriarchs are in many respects analogous to the imperial documents and symbolize the high status of the patriarch of Constantinople. They were, however, sealed with lead bulls. Byzantine private documents are almost exclusively notarial instruments. They are immediately recognizable by the crosses marked at the top of the documents. Used in lieu of the signature, the cross was the mark of the sender and contained his name and official function in one of its angles. The document was either signed by witnesses, or at least the cross preceding their names is autograph. Following that is the signature of the notary. These documents were not usually sealed.

The papal chancery. Knowledge about early papal documents is scant because no originals survive from before the 9th century, and extant copies of earlier documents are often much abridged. But it is clear that the popes at first imitated the form of the letters of the Roman emperors. The papal protocol consisted only of the superscription and address and the final protocol of the pope's personal "signature"—not a mention of his name but merely a blessing. Toward the end of the 8th century, it became customary in certain documents to mention in the final clauses the name of the scribe responsible for the drawing up of the document; this was given with the date of issue, indicated by month and indiction, immediately following the subject matter of the document. There followed another clause, the great dating formula, *datum per manus* ("given by the hand of . . ."), naming a high chancery official and giving the date by reference to the regnal years of both emperor and pope. Both were used in documents containing decrees of permanent legal force, which came to be called privileges. Under Pope Leo IX (1049–54), the benediction written by the pope was changed into a monogram not written by him, but his signature was now introduced, placed in a round symbol, the *rota.* By the early 13th century, papal documents had evolved into two distinctive groups: solemn privileges and letters. Solemn privileges can be distinguished by their enlarged letters (*elongata*) of the first line, by the phrase *in perpetuum* ("in perpetuity") at the end of the address, by a threefold *amen* at the end of the text, by use of the *rota,* the pope's signature, the monogram, signatures of the cardinals, and by the *datum per manus.* Among letters, those whose bull was fastened on silken cords (*litterae cum serico*) brought some benefit to the recipient, while those with bulls fastened on a hempen cord (*litterae cum filo canapis*) contained either orders or the papal delegation in a dispute.

Great dating formula

The number of solemn privileges began to decline from the mid-13th century, and eventually they were completely discontinued, their function being partly taken over by the *litterae cum serico,* which became increasingly elaborate in form. A new type of document also developed, the papal bull, distinguishable primarily by its use of formulas such as *ad perpetuam rei memoriam* ("that the matter may be perpetually known") in the superscription. Yet another new papal document appeared at the end of the 14th century, the brief (*breve*), used for the popes' private or even secret correspondence. Written not in the chancery but, instead, by papal secretaries (an office dating from about 1338), the briefs were sealed on wax with the imprint of the papal signet ring.

Chancery officials

The papal chancery of the 4th to the 8th centuries was similar to the late-Roman imperial chancery. Its notaries (*notarii, scriniarii*), organized in a guild (*schola*), were headed by the *primicerius notariorum* and the *se-*

cundicerius notariorum (first and second of notaries) and included the especially important notaries of Rome's seven ecclesiastical regions. But, during the 9th century, the *bibliothecarius*, the papal librarian, became the most important chancery official; a little later, various important bishops and dignitaries seem to have acted occasionally as *datarius* (the official named in the *datum per manus* formula). During the mid-11th century, a phase of German influence led to the temporary employment of notaries from the court of the emperor Henry III, who drew up papal privileges according to imperial formulas. A more important and permanent outcome of German influence was the gradual replacement of the *bibliothecarius* by a chancellor as the highest chancery official. The chancellor was invariably a cardinal, and in his absence another cardinal acted in his place as vice chancellor. Lesser chancery personnel still included the seven regional notaries; increasing business involved the use of lesser paid scribes in addition to the established notaries. From the late 11th century, a papal chapel, modelled on those of contemporary emperors and kings, developed, and its staff was often employed in chancery tasks.

From the early 13th century, the vice chancellor became the permanent head of the chancery, the office of chancellor remaining vacant. During that century the vice chancellors were ordinary clerics, who renounced the office if elevated to the cardinalate; thus, the chancery became directly subordinate to the pope himself. Both the numbers and the official standing of the notaries in the chancery, which then functioned entirely separately from the chapel, gradually increased. Higher chancery officials were often distinguished canonists (legal experts), such as Sinibaldo Fieschi (later Pope Innocent IV), Godfrey of Trani, and Richard of Siena. From the beginning of the 14th century, bishops or cardinals filled the office of vice chancellor. During the Great Schism (1378–1417) there were two papal chanceries and two vice chancellors, one in Rome and one in Avignon.

Under Innocent III the procedure of the papal chancery had changed. Letters concerning matters of import to the papal Curia (*de Curia*) were drafted by the pope himself or else by a cardinal, the vice chancellor, or a notary. But the majority of the papal documents were elicited by their recipients, who had first to present to a notary the substance of their petition in a form the text of which largely anticipated the wording of the desired document. Professional proctors attached to the Curia assisted in the drafting and were also responsible for the documents during later stages of the procedure. Once a petition was approved, the notaries or the *abbreviatores* drafted a suitable document, drawing on a selection of formula books. After a final copy (engrossment) had been made and checked, it was read, if necessary, to the pope or in a special department of the chancery, the *Audientia litterarum contradictarum*. It was then passed to the Cistercian lay brothers who had charge of the papal bull, sealed, and given to the petitioner, who had had to pay a fee at almost every stage of the proceedings.

Insufficient research has so far been done on the papal chancery during the 14th and 15th centuries. Whereas formerly, when the vice chancellor was absent, one of the notaries had deputized for him, a new official, the *regens cancellariam*, was now created to fulfill this function. The number of notaries increased steadily, and, from the 13th century onward, an increasing number of public notaries worked in the papal administration. In order to distinguish between them and the papal notaries proper, the latter became unofficially known as protonotaries. The notaries were now in charge of the letters of justice, while the letters of grace were handled by the *abbreviatores*. The scribes remained in charge of the engrossments. A *computator*, aided by several assistants, was responsible for collecting fees.

The royal chanceries of medieval western Europe. Of the nations that held power in western Europe after the collapse of the Roman Empire there, the Ostrogoths, who occupied Italy from the late 5th to the mid-6th century, took over the ancient Roman imperial-chancery system in its entirety. Very little is known about the royal documents of the Lombards, their successors in Northern Italy, since not one of them has been preserved in its original form. But Lombard officials in charge of drawing up the documents were still trained in the Roman tradition. As well as *referendarii*, there were notaries who also acted as scribes. It is very likely that all of them were laymen.

Until the 12th century two main types of documents, diplomas and mandates, were produced north of the Alps, in the Merovingian, Carolingian, German, and French royal chanceries. Very little is known about the Merovingian royal chancery and its organization. The names of the scribes are never mentioned in the documents, but they were signed by high chancery officials, the *referendarii*.

When the Merovingian dynasty was supplanted by the Carolingians, chancery procedure changed drastically. In contrast to the Merovingian kings, the first Carolingian king, Pepin the Short, was unable either to read or write. He therefore entrusted the responsibility for the correctness of the royal documents to an official of the court. At about the same time, the task of drawing up documents was taken over by those clerics whose original duty had been to look after the most important relic of the royal court, the coat (*cappa*) of St. Martin of Tours. Collectively named the *capella* (chapel), these clerks were individually called *capellani*, chaplains. This close connection between the court chapel and the chancery existed under the later Carolingians and at the German and French and other royal courts, including that of England. Until well into the 12th century, European chanceries were not bureaucratic offices in the modern sense but, rather, in most cases an assemblage of chaplains suited for the task of issuing documents and usually working under a cleric who was not the head of the chapel. Not all chaplains wrote documents, however, and the chapel and chancery thus remained separate institutions. From the reign of the emperor Louis I the Pious (814–840), the heads of the chancery were not personally involved in writing the documents, a task performed by unnamed and unknown scribes. At first the scribes were indiscriminately designated as either *notarii* or *cancellarii* (higher, Roman provincial officials of the 5th and 6th centuries, who stood at the barriers, *cancelli*, of the council rooms), but, by the 9th century, the title of *cancellarius* was gaining ground and was increasingly applied to the head of the chancery. The 9th century was a period of transition, during which, for a while, the archchaplain, the head of the chapel, became also the head of the clerks who wrote the charters.

The *capella* of St. Martin

Under the Ottonian dynasty, which came to power in the eastern division of the original Carolingian empire early in the 10th century, the German royal chancery developed the organization that was to characterize it throughout the remainder of the Middle Ages. The heads of the chancery were the archchancellors, but the office was entirely honorary and soon came to be automatically held, as far as Germany was concerned, by whoever was archbishop of Mainz. When the German kings or emperors established administrations in Italy, Italian bishops were at first made archchancellors for Italy, but in 1031 the office was attached to the archbishopric of Cologne. From the 11th century, Burgundian bishops were archchancellors for Burgundy, but, in the second half of the 13th century, the archbishop of Trier took over the office.

Imperial chancellors

The actual heads of the chancery were the chancellors. At first there was a chancellor, as well as an archchancellor, for each separate part of the empire—Germany, Italy, and Burgundy—but from 1118 there was only one chancellor for all three kingdoms. But even the chancellors, all of whom were clerics, were rarely involved in the actual composition and engrossing of documents, being usually engaged, as important advisers to the king or emperor, in much weightier matters. They do seem to have been especially concerned, however, with decisions about the granting of charters, and they supervised the work of the scribes or notaries. From among the ranks of these notaries, a group of protonotaries gradually developed after the mid-12th century, as a result of influence from the chancery of the Norman rulers of Sicily. Often called upon to deputize for the chancellor, the protonotaries, from the

late 13th century onward, frequently titled themselves vice chancellors.

From the 12th century onward, the documents issued by the German royal chancery were divided into various classifications. The diploma, by then usually called a privilege, existed in two categories, the solemn and the simple privilege. A solemn privilege included the *invocatio,* the signum and recognition line, and a detailed dating or at least one of these three elements, which were entirely lacking in simple privileges. Gradually, simple privileges merged into documents called mandates; it is not always easy to distinguish between them, but, in general, privileges were concerned with rights in perpetuity, while the mandates dealt mainly with matters of only temporary importance. From the early 14th century, mandates were superseded by the use of letters patent and letters close (open or closed letters). Privileges continued to be sealed with a hanging seal; the seal on letters patent was impressed on the document and was used to seal up letters close.

As the power of the German kings declined during the later Middle Ages, so that of the archchancellors increased, and in the 14th century they attempted to win control of the chancery. But, despite fluctuations in the power struggle, the king retained control of the chancellor, who, by the end of the 15th century, held the title of imperial vice chancellor.

Under the Carolingians and the first Capetians in France, various bishops and archbishops, especially the archbishops of Reims, held the office of royal chancellor. But at that time the office was merely titular, and, by the end of the 11th century, it disappeared entirely. From the 12th century onward, the title of chancellor became reserved to the head of the chancery. These new chancellors became so powerful that in 1185 King Philip II Augustus left the office vacant, and, during almost the whole of the 13th century, the chancery was administered by subordinate officials. Chancellors, often laymen, were appointed again in the 14th century, however, and the office remained important until 1789. As in other parts of Europe, the French chancellor merely directed the work of the notaries, and it was they who were responsible for drawing up the documents. From 1350 onward, the notaries were called secretaries, and both their numbers and their importance steadily increased. From the 15th century, the tremendous expansion of business occupying the Grande Chancellerie led to the establishment of several subsidiary *petites chancelleries,* all issuing royal documents sealed with the king's signet. Until the reign of Henry I (1031–60), the old Frankish type of diploma was issued almost exclusively. Then, gradually, charters in the simpler form of letters began to replace the diplomas, and, during the 13th century, the *lettres patentes* became the common type of document. These lacked the *invocatio,* the monogram, and the signature of the high dignitaries, and they gave the simple form of dating. From the 14th century, two forms of *lettres patentes* existed: the *charte,* which was sealed with a green wax seal hanging on red and green silk cords; and the *lettre patente,* used mainly for administrative mandates, which was sealed with a yellow wax seal on double and single cord. Besides, *lettres closes* were used from the 13th century onward.

The English royal documents of the Anglo-Saxon period (before 1066) can be divided into two large groups: the charters, mostly written in Latin; and the writs, written in Old English. The charters, for the most part concerning grants of land, began with either a verbal or a symbolic *invocatio* (a cross or the monogram XP for Christ). There was an *arenga* but no *intitulatio.* Charters were not sealed, the validation comprising the nonautographed signatures of the king and of ecclesiastic and secular dignitaries. Many of the extant charters of this era are forged or interpolated; the series of those apparently genuine starts shortly after the arrival in Canterbury (669) of the archbishop Theodore, and these show similarities with late-Roman private documents. It therefore seems probable that the Anglo-Saxon charters derive from Italian models brought to England by the Roman missionaries. Most of the charters were apparently written by the recipients. From the 11th century, the charters were gradually superseded by sealed writs, which became the most important type of document in medieval England. At first written in Anglo-Saxon, they were produced in Latin soon after the Norman Conquest. No continental document was anything like them. Written on a narrow strip of parchment, their entire text occupied only a few lines. A symbolic *invocatio* in the form of a cross was followed by the king's name and the address, which contained a salutation clause. There was no *arenga,* and the address was followed by a short description of the bequest, if the writ concerned a grant, or a directive, if the writ was a mandate. The Anglo-Saxon writs had no signatures of witnesses and no date, but after 1066 these elements were added. The dating consisted at first only of a mention of the place of issue, days and month being usually given only toward the end of the 12th century. From the very beginning the writs were supplied with a pendant seal as a means of validation. From the reign of Henry II (1154–89), it is possible to distinguish mandate-like writs of the old type and charter writs, which mainly concerned questions of feudal enfeoffments and confirmations of privileges and were usually more carefully executed than the mandate writs. By the early 13th century, the charter writs had developed into a new form of charter. It contained the *intitulatio,* including all the titles of the king; the general address; the dispositive text, which in most instances was introduced with the expression *sciatis* ("know that"); and the final clauses, consisting of the names of several witnesses and of the *datum per manus* clause mentioning the chancellor and giving the date and place of issue. During the final stage of its development, the great seal of green wax, pendant on silk cords, served as the means of validation. The ordinary writs further evolved into the common-law writs (containing orders to persons mentioned by name), the writs of summons, and, particularly, the letters patent (the latter eventually assuming the function of the charter writs, which disappeared at the end of the 13th century) and the letters close. The letters patent were furnished with a general address. Often introduced with the *sciatis,* the text concerns either limited grants or commissions to royal officials, introduced by the word *precipio* ("I order") or a similar expression.

English writs

The dating begins with a series of witnesses and contains the place of issue, day, month, and year of the king's reign. Occasionally the dating is followed by the words "per N.," which are assumed to designate the household official transmitting to the scribe the royal order to issue the writ. The means of validation was the great seal in white wax, pendant on a single strip of parchment (*simple-queue*) or on a double strip of parchment or silk (*double-queue*). The letters close obtained their name from the fact that they were closed by means of the great seal. They contained either commands or information directed to a single individual or to several persons. Characteristic of these letters are the words *teste me ipso* ("witnessed myself") introducing the regular dating clause. These types of royal documents remained essentially unchanged throughout the Middle Ages and were imitated by both secular and ecclesiastic English magnates and dignitaries.

The English royal chancery grew out of the royal household. As on the Continent, it was at first merely a group of royal chaplains, and, until the Norman Conquest, there was no chancellor at their head. The chancellor was the keeper of the seal but usually took no part in the issuing of documents. During the 12th century the number of scribes was still low, between two and eight. At first this "chancery" travelled about with the king, only in the course of the 13th century establishing a permanent location at Westminster. At first English royal documents were not dispositive but merely evidentiary, confirming a previously arranged, orally discharged legal act, and the king's order for issuing the document was given orally. But, from the late 13th century, a royal secretariat came into existence, which was in charge of relaying to the chancellor, by warrant sealed with the privy seal, the royal order for issuing a writ under the great seal. In many ways the secretariat thus became a competitor of the chancery. In addition, during the first half of the 14th century, the Signet Office was established, so called after the small seal (signet). The king's secretary was also the head of this office. All these

The Signet Office

shifts made the issuing of royal documents increasingly complicated. From the end of the 14th century, the common procedure involved, first, the petitioner submitting a petition to the king. If the king approved of it, his secretary forwarded a warrant carrying the signet to the keeper of the privy seal with the request to send, in his turn, a warrant carrying the privy seal to the chancellor. The chancellor then ordered the issue of the document, which would bear the great seal. So far as the issuing of royal documents is concerned, the fact that the secretariat developed into the secretariat of state is of great significance. The king's secretary (later on, there were two of them) became the centre of the royal administration. It was in his office that the state papers originated, which already under Henry VIII (died 1547) far surpassed in importance the old chancery records. Among the state papers there are in-letters, out-letters, drafts, reports, and schedules. The decline of the rolls (document registers) during the 16th century gave rise to yet another new office, the State Paper Office, headed since 1578 by the clerk of the papers. The second holder of this office, Sir Thomas Wilson, established the division of the state papers into foreign and domestic. As departments of state proliferated during the 18th and 19th centuries, they developed their own archives. In 1838 all the public archives became subject to an official called the master of the rolls.

Smaller European nations usually modelled their documents on those of the papacy, the empire, France, or England. The influence of papal letters and privileges can be observed particularly in Aragon, Castile, and Portugal, while German royal diplomas served as models in Bohemia (which was part of the empire), Hungary, and Poland. Because of the close political ties between the two kingdoms, Anglo-Saxon influence can be traced in the seal of the royal Danish documents during the 11th century, but, in the course of the 12th century, royal and princely German documents became the models for Danish as well as Swedish royal documents. Norwegian royal documents were modelled on Anglo-Saxon writs, probably as a result of the influence of English missionaries working in Norway from the early 11th century. The Norwegian writs were drawn up in the vernacular. The chancery of the Norman rulers of southern Italy and Sicily was highly developed. Influenced by the form of papal documents, the Norman documents comprise mainly privileges, either formal (with *rota*, witnesses, and gold bull) or simple (with *rota*, leaden bull, or wax seal), and mandates. They all have a detailed dateline that includes the name of the chancellor or other high court officials, the number of years since the birth of Christ, the regnal year of the king, and the *apprecatio*. The mandates are more simply executed. They lack the *invocatio* and start out with a simple *intitulatio* and *inscriptio* that ends with a salutation clause. There is neither *arenga* nor *corroboratio*, but there is a command clause in the text. The dating consists only of the place of issue, the day of the month, and the *indictio*. There is no *rota* or signature. The seal is of red wax. The head of the royal chancery of the Norman kingdom of Sicily was the chancellor, a layman who was an influential court official. The notaries who drafted and wrote the documents were also laymen. Because the German Hohenstaufen emperors also ruled in Sicily from 1194 to 1250, Norman chancery practice influenced subsequent German documents. (Pe.He.)

The chancery of Norman rulers

EPIGRAPHY

Epigraphy (from the classical Greek *epigraphein*, "to write upon, incise," and *epigraphē*, "inscription") denotes a branch of scholarship devoted to the study of written matter recorded on hard or durable material. Because such media were exclusive or predominant in many of mankind's earliest civilizations, epigraphy is a prime tool in recovering much of the firsthand record of antiquity. It is thus an essential adjunct of the study of ancient man; it secures and delivers the primary data on which historical and philological disciplines alike depend for their understanding of the recorded past. In a narrower sense, epigraphy is the study of such documents as remains of the written self-expression of early cultures and as communication media in their own right, attesting to the development of visible sign systems and the art of writing as such. Finally, in later periods including our own, where perishable writing media predominate, epigraphy affords insights into the styles and purposes of monumental or otherwise exceptional techniques of written recording.

Uses of epigraphy

Materials and techniques. The delimitation of epigraphy vis-à-vis contiguous and related areas of antiquarian scholarship meets with some ambiguity. In a wide sense, epigraphy concerns itself with the total firsthand transmission of the written remains of ancient civilizations (as opposed to *post-factum* copying) and the nature of the material (*e.g.*, stone, marble, metal, clay, terra-cotta, pottery, wood, wax tablets, papyrus, parchment) and the technique of recording (cutting, carving, engraving, casting, embossing, scratching, painting, drawing, etc.) have mere secondary relevance. Under this maximum definition certain subdisciplines may be included under the overall canopy of epigraphy: notably numismatics, which concerns itself with legends on coins and medals, and papyrology, the study of a special type of perishable record that is normally preserved only in the dry climate of Egypt and in adjacent desert regions. In the case of Egypt, papyrology tends to impinge upon wood and clay media as well, thus leaving mainly stone and metal objects as the concern of epigraphy proper.

In general, however, unless so subdivided, epigraphy encompasses inscriptions at large, be they on primary writing surfaces or on such assorted objects as vases, potsherds, gems, seals, stamps, weights, rings, lamps, and mirrors. A further related discipline is paleography, which concerns itself with the study of scribal hands and styles of writing and has significance for the dating of epigraphic as well as other written documents.

The nature of the materials and techniques used for inscriptions is closely tied to the external purpose of the record itself. Thus, inscriptions may be divided into monumental, archival, and incidental. Monumental inscriptions were intended for enduring display and were therefore, as a rule, executed in lasting material, such as stone or metal. Maximal exposure to mortal eyes need not have been the prime purpose of their originators—*e.g.*, the tomb chambers of Egyptian pharaohs, intended to be sealed forever, had their inner surfaces covered with monumental hieroglyphs; the great Bisitūn inscription of King Darius I of Persia is on a high rock surface and legible only after precarious rock-climbing or from airborne conveyances. Under this classification may be included also micromonumental inscriptions found on such objects as coins, seals, and rings, meant to endure in their own right.

Monumental, archival, and incidental inscriptions

Archival inscriptions were essentially a feature of those early societies that kept records and that used such materials as have been preserved thanks to their intrinsic, accidental, or incidental durability. Many ancient Near Eastern cultures employed clay tablets for writing, which they fired to insure their soundness. Minoan and Mycenaean archivists in ancient Crete and Greece used perishable temporary clay records that were preserved by unintentional baking in the conflagrations that destroyed their storerooms. Papyrus records from Egypt have survived as a result of climatological chance—mainly low humidity. The official purposes of public display and of archival preservation were sometimes complementary, and therefore coincidental or overlapping matter has been preserved. In some cultures the techniques employed in monumental and archival writing tended to differ (notably in Egypt, where increasingly cursive hieratic or demotic script contrasted sharply with monumental hieroglyphic), and occasionally the language itself would be different (for example, in the Hittite Empire, where the clay tablets in cuneiform employed mainly straight Hittite or Akkadian, whereas the monumental "hieroglyphic" rock inscriptions and seals used a distinct language).

Incidental inscriptions may be defined as those not seriously meant for preservation. They include, for example, wall scrawlings of the graffiti type and casual records that were kept on cheap writing matter such as potsherds (*ostraca*) and scraps of papyrus. Many a city dump of ancient Egypt has yielded a rich harvest for the study of daily life.

Inscriptions as historical source material. In study-

Evidence of extinct civilizations

ing the political, administrative, legislative, and dynastic records of extinct civilizations, the modern historian must bring to bear all the evidence at his disposal; and such evidence may vary sharply from one locality and period to another. Historiography in the modern sense—the analytical ordering and interpretation of past institutions and events—is an invention of ancient Greece, and even there it only gradually eschewed the fabulous. In many early societies (*e.g.*, Mesopotamia, Egypt), chronographic records were either annalistic or legendary in kind and sometimes apologistic or propagandistic in purport; and in others (*e.g.*, India) a cyclical world view prevailed, and a lack of feeling for linear time depth precluded an ordered appreciation of the past, leaving only legendry—in such cases only synchronisms with time-anchored events elsewhere allow a precise dating.

Thus, the amount of predigested ancient information bearing on antecedent events may vary from sophisticated literary to scrupulous epigraphic or may be wholly lacking or largely valueless. In the latter instances the historian is almost exclusively dependent for native information on primary documents, and such documents are in most cases inscriptional.

Ancient Mesopotamia. Surviving epigraphic matter from the 3rd and early 2nd millennia BC includes both historical and quasi-historical material. The Sumerian king list is a compilation of names, places, and wholly fabulous dates and exploits, apparently edited to show and promote time-hallowed oneness of kingship in the face of the splintered city-states of the period. The Sargon Chronicle is a piece of literary legendry concentrating on spectacular figures and feats of the past, whereas contemporary royal inscriptions, notably by Sargon I of Akkad and Gudea of Lagash, are historical documents in the proper sense.

Both kinds of texts are preserved also from the Babylonian and Assyrian periods, from the reign of Hammurabi (1792–1750 BC) to the 6th century BC. There are lists of date formulas and year names from Hammurabi's reign and from that of his son Samsuiluna; lists of Assyrian eponymous year names, based on those of dignitaries; the Babylonian king lists, running from Hammurabi through the Kassite era and the Assyrian domination of Babylon to the last flicker of Babylonian self-assertion in the early 6th century BC; the Assyrian king list from Khorsabad, which made good use of earlier compilations; and notably the so-called Synchronistic Chronicle, which juxtaposed the kings of Assyria and Babylonia in the same millennial sequence. Historical documents comprise, above all, the stately sequence of annals by the kings of Assyria, recorded on stone slabs, stelae, foundation markers of buildings, bronze gates, statues, and obelisks and in clay archives (prisms, cylinders, tablets). Starting in the Old Assyrian period, they were especially extensive in the reigns of Tiglath-pileser I (1115–1077 BC), Ashurnasirpal II (883–859 BC), Shalmaneser III (858–824 BC), Adadnirari III (810–783 BC), Tiglath-pileser III (744–727 BC), Shalmaneser V (726–722 BC), Sargon II (721–705 BC), Sennacherib (704–681 BC), Esarhaddon (*c.* 680–669 BC), and Ashurbanipal (668–627 BC).

For all their swaggering bombast and flaunting of deliberate cruelty, the annals provide prime historical source material. The detail of the Assyrian conquest of Syria, Palestine, parts of Asia Minor, Cyprus, Arabia, and Egypt would be spotty indeed without recourse to these annals, for they show the centre of political power, unlike such provincial records as those from contemporary Egypt or the Old Testament.

Legal compilations and law codes

Legal compilations and law codes also have pride of place in the epigraphic record of ancient Mesopotamia. These form a unique succession, starting in the 3rd millennium BC with that of King Ur-Nammu of the Sumerian 3rd dynasty of Ur (*c.* 2100 BC), continuing with those of the Sumero-Akkadian king Lipit-Ishtar (in Sumerian) and King Bilalama of Eshnunna (in Akkadian) during the interval of the 3rd dynasty of Ur, and the rise of the Amorite dynasty of Hammurabi (*c.* 2000 BC), culminating in the great diorite stela of Hammurabi (*c.* 1750 BC), showing retardation and recrudescence in the Middle Assyrian laws that are found on clay tablets at Ashur (at the time of Tiglath-pileser I), and petering out in the fragmentary Neo-Babylonian laws dating from the 7th century BC.

The stela of Hammurabi must have been originally set up in some Babylonian population centre for the literate to read and know their rights. Some Elamite invader must have carried it off to Susa (perhaps *c.* 1200 BC), where it was found in 1901 and removed to the Louvre in Paris. The bulk of the stela contains the text of the code, partly erased on the obverse but restorable in some measure from clay-tablet versions of the same laws. The top depicts the king in a worshipful pose, receiving the laws from the sun god, Shamash. In reality Hammurabi—the sixth of 11 kings of the Old Babylonian or Amorite dynasty—was a practical codifier rather than a revelatory mediator of law. His code was an effort to fuse into a workable whole the ancient inheritance of Sumerian-based jurisprudence and the Semitic talion law (punishment according to the "eye-for-an-eye and tooth-for-a-tooth" principle) of the Akkadian superstratum. The result is not a model of economy or arrangement or logical organization, but the code of Hammurabi constitutes nevertheless the first great legal monument in the history of mankind. The later Assyrian laws show traces of further removal from the cradle of Sumerian civilization since they are both harsher and noticeably more primitive.

Ancient Egypt. Egypt attracted the special curiosity of the Greeks, and Herodotus (5th century BC) devoted an entire book to on-the-spot observations and fanciful tales about the land of the Nile. The lost *Aigyptiaka* (or *Aegyptiaca*) of Manetho (3rd century BC) contained the roster of 30 dynasties, which still underlies the chronology of ancient Egypt. Such classical writers as Strabo, Plutarch, and Pliny the Elder all dealt with various aspects of Egyptian antiquities.

Yet the fund of knowledge would be woefully skeletal and inaccurate without the explicit testimony of contemporary records from Egypt itself. The decipherment of the Egyptian writings gave the impetus to Egyptian epigraphy. The progress of excavations multiplied the corpora of texts, especially adding the papyrological dimension. In addition, cuneiform Akkadian on clay tablets was the international diplomatic medium of writing during the most brilliant phases of Egyptian history and is hence an integral part of the Egyptian epigraphic record.

Decipherment of Egyptian writings

The historically significant Egyptian epigraphic texts, apart from their external peculiarities, have likewise special traits relating to genres. There is little attempt at historiography and great fluctuation in bulk in the course of dynastic vicissitudes. They are mainly annalistic and thus firsthand accounts of pharaonic or other high-level deeds; but the peculiar features of stylization, stereotyping, and usurpation must frequently give the careful historian pause and sometimes debase the face value of the record. Written monuments became somewhat numerous during the 4th dynasty (*c.* 2575–*c.* 2465 BC), that of the pyramid builders Khufu, Khafre, and Menkaure. Notable are the fragmentary annals of Snefru, which already alluded to dealings with Asia, the chronic goal of Egyptian territorial ambitions. Historic records persisted under the following two dynasties, with particular articulateness in the reign of Pepi I, third king of the 6th dynasty (*c.* 2325–*c.* 2150 BC), then subsided until a modest reemergence in the (Theban) Middle Kingdom of the 12th dynasty (1938–*c.* 1756 BC). Another silence shrouded the period of the Hyksos kings (*c.* 1630–*c.* 1523 BC), broken only subsequently by such retrospective revulsion at the memory of barbarian domination as that in Queen Hatshepsut's (1479–58 BC) temple inscription at Istabl Antar in Middle Egypt. The golden age of historical recording began in the 15th century BC with the central rulers of the 18th dynasty, notably Thutmose III and Amenhotep II and III. Thutmose's annals on the walls of the temple of Karnak describe 20 years of ceaseless military activity in Asia, some 16 campaigns in all, and are supplemented by stelae from Armant in Upper Egypt and Gebel Barkal near the Fourth Cataract, as well as by lists of conquered lands at Karnak. Similar material continued in the reigns of Amenhotep II and III, in the latter's case importantly supplemented by the cuneiform correspondence with foreign powers (Mitanni, Arzawa,

Golden age of Egyptian historical recording

etc.), which was subsequently stockpiled and archived by Ikhnaton in his transitory new capital, where it lay buried to await the modern excavators of Tell el-Amarna. Ikhnaton's religious preoccupations (he changed the official religion to the worship of the sun god Aton), and political apathy led to the loss of many of Egypt's Asian possessions. Records of Ikhnaton's short-lived son-in-law, Tutankhamen, at Thebes (1332–23 BC), make recantation and restoration for the heresy. Tutankhamen's successor, the warlord-pharaoh Horemheb, left boastful accounts of foreign conquest that sound suspiciously grandiose in relation to plausible reality.

In the 19th dynasty (1292–1190 BC) Seti I went to war against the Syrians, Hittites, and Libyans, letting the world know about it on the walls of Karnak. But in this respect he was no match for his long-lived son, Ramses II, who usurped the monuments of others and covered unprecedented amounts of wall space with his own real or inflated exploits. (The Battle of Kadesh against the Hittites in 1299 BC, which ended in a stalemate, was given lavish coverage as a triumph on temple walls at Karnak, Abydos, and Abu Simbel.) In the 20th dynasty (1190–1075 BC) occurred incursions of the "sea peoples," and the records of Ramses III detailed both the crisis and the increasing accumulation of wealth and power in the religious establishment. Subsequently, Egyptian history receded from the world scene, with "Libyan" and "Ethiopic" dynasties and a brief Saite renaissance of the 26th dynasty (664–525 BC), already under the Assyrian and Babylonian shadow, soon to be replaced by the Persian. The firsthand political records declined accordingly, although they remain of significance for local history down to the Ptolemaic era, a dynasty that ruled Egypt beginning in 304 BC, founded by Ptolemy I Soter, a general under Alexander the Great.

No law codes have been found in Egypt, presumably because codification was not practiced. There are, however, royal administrative and legal decrees granting privileges and immunities and also records of legal proceedings, especially of the Theban tomb-robbery trials during the 20th dynasty.

Other ancient Middle Eastern regions. Regions adjacent to the power centres of Mesopotamia, Egypt, and Iran were frequently mere political and administrative adjuncts, often obscure vassaldoms or adversaries without notable or attested written traditions. The Mitanni kingdom in northern Mesopotamia had some ephemeral big-power dealings with Egypt in the days of Amenhotep III, but its capital city is still lost in the sands, and thus its presently known epigraphic tradition is merely part of the correspondence in the Tell el-Amarna archives. The records of the Elamite kingdom with its capital at Susa were mostly ancillary to Mesopotamia in the 2nd millennium BC and to Iran later on. The region of Syria and what later came to be called Palestine was in the 2nd millennium the object of an extended tug-of-war between Egypt and the Hittite kingdom.

The Hittites were, in fact, the third great international power in the Near East during part of the 2nd millennium BC, and the epigraphic yield of their royal archives at Boğazköy in central Asia Minor matches or even surpasses in richness that of Mesopotamia and Egypt for the few centuries in question. The cuneiform records of the Hittites contain a tradition of unique royal political self-expression. These documents begin with the oldest known Hittite text, the inscription of the early ruler Anittas, detailing dynastic struggles of an obscure and possibly apocryphal past. From the founder of the Old Kingdom, the firmly historical Hattusilis I (Labarnas II), came an annalistic autobiography (excavated in 1957) and a "farewell address," or political testament, in Hittite as well as Akkadian versions. Subsequent events, including the capture of Babylon by Hattusilis' son, Mursilis I (*c.* 1590 BC), and the succeeding era of regicidal upheavals, are known from an edict of King Telipinus, who detailed them as he set about regulating the rights of royal succession. The subsequent founder of the Hittite Empire, Suppiluliumas I (*c.* 1350 BC), and his son Mursilis II left annals detailing their military and political deeds. Mursilis was a particularly prolific annalist and edited his father's annals as well.

The earliest Hittite texts

The great encounter with Ramses II at Kadesh in 1299 BC occurred in the reign of Mursilis' son, Muwatallis, and left an echo in the autobiography of his brother and successor, Hattusilis III. Hattusilis' autobiography is a tract of self-justification for a breach of the edict of Telipinus in deposing his nephew and predecessor Urhi-Teshub (Mursilis III).

Other Hittite documents inveigh against treasonable behaviour ("Indictment of Madduwattas") or contain detailed instructions for military, civil, and court officialdom. Hittite queens had prerogatives of independent high-level initiative, and examples of their correspondence with foreign potentates supplement the archives of their husbands. The most remarkable external political documents are numerous state treaties, sometimes between equals but more often covenants specifying protectorate or vassaldom status for subordinate states on the fringes of the kingdom. Equally notable is the Hittite Law Code, relatively enlightened and mild in the face of its contemporary counterparts in Mesopotamia. Altogether, the inscriptional documents are practically the exclusive source material for knowledge of the Hittites; not even the existence or location of their empire was surmised prior to the discovery of their archives.

After the collapse of the Hittite Empire (*c.* 1190 BC), significant records from Asia Minor ceased for many centuries, whereas local history in the Syro-Palestinian area was recorded in the inscriptions of petty dynasts increasingly under the shadow of Assyrian domination. The break with the past is evident in the writing systems (Hittite Hieroglyphs or West Semitic alphabet rather than cuneiform) and in the languages (Indo-European Anatolian, Canaanite, Aramaic). Into this category fall the stela of King Mesha of Moab (*c.* 830 BC) now in the Louvre, the Phoenician-Hieroglyphic Luvian bilingual inscription of Azitawadda of Adana (late 8th century BC), and those of the kings of Ya'diya-Sam'al. Contemporary cuneiform documents from the Urartu kingdom around Lake Van in eastern Anatolia are historically and culturally an offshoot of the history of 8th-century Assyria.

Ancient Iran. Epigraphically recorded history in ancient Persia began dramatically with the rise of the Achaemenid dynasty in the 6th century BC. Cyrus II the Great's conquest of Media, Lydia, and Babylonia, Cambyses' occupation of Egypt, and the incursions into Greece of the succeeding side branch of the family, beginning with Darius I, created in short order a world power destined for centre stage on the international scene for the following two centuries. The international character of the empire is reflected in the frequently trilingual royal inscriptions—with Akkadian and Elamite versions in traditional syllabic cuneiform, and the Old Persian text in its own simplified quasi-alphabetic system of wedge-shaped writing. The Achaemenids' time span ranges from Darius' great-grandfather, Ariaramnes, to their less glorious progeny, who were ultimately extinguished by Alexander the Great. The empire was centred in Persia; but a granite stela of Darius, found near the Suez Canal, recorded in Old Persian, Elamite, Akkadian, and hieroglyphic Egyptian the opening of a canal from the Red Sea to the Nile. The epigraphic material included rock surfaces, building walls, columns, doorways, cornices, statues, and doorknobs; bricks, plaques, plates, and tablets of clay, stone, gold, and silver; vases; weights; and seals. Almost all the longer texts were by Darius and Xerxes I; an important one of Xerxes was found on a stone tablet at Persepolis in 1967. The great Bisitūn rock inscription of Darius runs to several hundred long lines in Old Persian alone, besides the Elamite and Akkadian versions. It is accompanied by 11 minor inscriptions, serving as keys to the sculpted scene of the panel, which shows Darius triumphant over the usurping impostor Gaumata and nine other rebels. The text is a self-statement of how Darius gained and consolidated his rule. It apparently had currency in the realm, apart from being tucked away on a sheer cliff wall, for a partial duplicate of the Akkadian version has been found on a dolerite (basalt) block from Babylon, and papyrus fragments from Elephantine have yielded scraps of an Aramaic edition. One peculiar documentary value of the text is that the

Rise of the Achaemenid dynasty

same period in Persian history is extensively covered in the Greek literary tradition by Herodotus, Ctesias, and others, and scholars can thus juxtapose Darius' own accounts with those of almost contemporary foreign historians. As an example, Darius stressed his role as saviour of the fatherland from the clutches of an upstart who pretended to be Smerdis, the brother of Darius' predecessor Cambyses. The latter had murdered Smerdis and was carrying on various outrages in Egypt when word came of the impostor's takeover back home. Darius stated that thereupon Cambyses "died his own death," meaning that it was a fatal matter without human interference, and that thus Darius' hands were clean in taking action against the impostor. According to Herodotus, Cambyses' legendary death involved a freak occurrence as he prepared to leave for home—an accidentally self-inflicted wound leading to gangrene. Both the meaning and intent of Darius' description are thus confirmed.

Juxtaposition of Darius' texts with Herodotus' historical writings

From later Arsacid (Parthian) and Sāsānid periods of Iranian history, there are likewise royal inscriptions that shed light on their respective eras down to the Islāmic conquest in the 8th century AD, and new specimens are still being discovered.

Ancient India. India's past became anchored in historical time and separable from legend only with the establishment of firm synchronisms with outside data. One such link is the Seleucid embassy of Megasthenes to the Maurya king Candragupta (Greek Sandrokottos) at Pāṭaliputra (Greek Palimbothra) in Magadha (modern day Bihar). The Maurya dynasty was continued in the early 3rd century BC by Candragupta's son Bindusāra (Amitrochates in the Greek sources) and had extended its power over much of the subcontinent. But then the Greek sources fall silent, and Indic literary tradition supplies only the usual web of timeless legendry. At this point, however, epigraphy makes a unique contribution in the form of the first authentic and datable historical documents from India, the edicts of Bindusāra's son and successor Aśoka. As a matter of epigraphic fact, Aśoka ruled all of northern India and a large portion of the south, from Taxila and beyond to Mysore and Kaliṅga (coast of Orissa and Andhra Pradesh). His 14 rock edicts and seven pillar edicts in numerous versions and copies, plus separate minor texts, are scattered over this expanse—in the Prākrit language of his time and in the Brāhmī script, except for some northwestern examples of the Aramaic-inspired Kharoṣṭhī writing. Even a Greek-Aramaic bilingual version was found in 1958 near Kandahār in Afghanistan. Aśoka's edicts are proclamations and ordinances in a Buddhistic spirit, designed to impart good order, morality, and moderation by the Emperor's personal enjoining and example. Particularly notable is the 13th rock edict, which bares the ruler's pangs of conscience over the conquest of Kaliṅga eight years after his coronation, his continuing sorrow over the cruelties committed, and his pledge to substitute the victory of Buddhist religious law (*dhamma*) for all earthly conquest.

Aśoka's edicts would rate a mere historiographic footnote for their inconsequential transitoriness, were it not that in the same breath Aśoka supplied the very synchronisms that are the main key to ancient Indian chronology. Among his western neighbours he mentioned Aṃtiyoge (Antiochus II Theos of Syria), Tulamaye (Ptolemy II Philadelphus of Egypt), Aṇtekine (Antigonus II Gonatas of Macedonia), Maka (Magas of Cyrene), and Alikasudaro (Alexander of Epirus or Alexander of Corinth). The dates of these contemporaries circumscribe the time of Aśoka's reign; combined with the earlier Greek synchronisms, they afford a firm foundation for the correlation of Indic and Mediterranean events. The edicts of Aśoka are thus a prime example of the value of inscriptions for historiographic dating and constitute a fixed record unparallelled in ancient Indian tradition. Later periods of Indian history, such as those of the Indo-Scythian and Gupta rulers, are also represented in epigraphic documents of some historical value.

The main key to Indian chronology

Ancient China. In China also, inscriptions are a means of separating chronological fact from historiographic legend. Nonepigraphic book composition on wood or bamboo strips had an early history in China, beginning in the later 2nd millennium BC; its scope was such that the Ch'in emperor Shih Huang-ti went down in history as a book burner in 213 BC. The San Tai, or three periods of early Chinese history (Hsia, *c.* 2205–1766 BC; Shang, *c.* 1766–1122 BC; Chou and Ch'in, *c.* 1122–206 BC), were long considered by Western scholars purely legendary down to the early Chou period, and the literary documents (such as the *Shu Ching* or "Classic of History") were dismissed as compilations consisting mostly of successive overlays of little historical value. But the historicity of written records from the later Shang era (*c.* 1400–1122 BC) is now apparent from the mass of inscribed archeological material found especially in northern Honan Province. These include, in particular, the so-called oracle bones (mostly tortoise shells and scapulae of animals), bearing incised records of royal divination. At the site of the Shang capital, Yin, were discovered inscribed vessels of bronze, bone, pottery, jade, and stone, probably ceremonial in nature and related to official ritual uses such as ancestor worship. The script is a mixture of pictograms, word signs, and phonograms. From the ensuing Chou era, bronze inscriptions of official provenance have likewise been found, especially records of royal largesse. Inscriptions from later periods form a steady but subsidiary source of information beside the larger, nonepigraphic written record.

Ancient Greece. The historically significant epigraphic record of classical Greece differs in many ways from most of those discussed above. Much of it is parallelled by a mature and independent tradition of professional literary historiography. Except for the preclassical Helladic (Mycenaean) period of the 2nd millennium BC (see below), there was no archival tradition, although the bulk of "monumental" records sometimes approximates the same purpose of massive preservation. There was no all-important power centre and no dominant rulership before Hellenistic and Roman times: thus the geographical scattering of records was extreme, although naturally with some focuses of emphasis such as Athens. Above all, there was continuity from the inception of literacy, with gradual but steady increase in bulk.

The special circumstances of the Greek epigraphic record

Epigraphically transmitted historiography in Greece is extremely scarce because the probing of past events has passed beyond the stage of dynastically centred and sheltered annalism; an example is the "Marmor Parium" (from the island of Paros and now at Oxford), which contains a chronographic rundown of traditional dates and events of Greek history. Rather than monolithic records of autocracy, there is history in the making by a plethora of tyrannical, oligarchic, or democratic microentities. Treaties of alliance and various other agreements between the multiple city-states—recorded on metal or stone and publicly displayed, or consecrated at such pan-Hellenic sanctuaries as Delphi or Olympia—form an important part of the epigraphic yield. Joint-citizenship covenants, decrees concerning the return of exiles, monetary agreements on coinage and debts, and tribute lists are typical examples. They are supplemented by the records of arbitration of interstate disputes, most often boundary matters, by third-party commissions. Thus, when around 240 BC a territorial disagreement arose between Epidaurus and Corinth, the Achaean League appointed a group of 151 Megarians as mediators, and their report survives. Further extensions of such "international" documentation are the *proxenia* decrees, which amount to letters of patent and resolutions of appreciation issued by one state to a citizen of another for service as *proxenos*, a kind of honorary consul looking after the interests of the other state's citizens. The extensive colonization efforts by the Greeks around the Mediterranean produced a further kind of political document—regulations governing conditions for emigration and return, citizenship rights of the colonists, and relations between the colony and the mother community. Not all historically meaningful international records are of the monumental type. Greek mercenaries of Pharaoh Psamtik II (ruled 594–589 BC) left their scrawlings on the legs of a colossal statue at Abu Simbel on the upper Nile, proving by their names and dialect that they came from Rhodes and Ionia and were far abroad on foreign adventure.

Internal documents of the various Greek states include numerous records of decrees and ordinances, both administrative and legislative. Stereotyped Athenian ones are complemented by variant forms in other localities; most contain a preamble setting forth the date and the officialdom in charge, the circumstances occasioning the action, the decision itself, means and sanctions for its enforcement, and sometimes instructions for providing and affixing the very physical record that has been preserved. Sometimes they amount to formal laws, such as those directed against extravagances in funeral practices.

Financial data of the states were minutely and permanently inscribed on stone, and the accounts thus displayed recorded in detail the receipts, expenditures, and balances of public funds. Very specific reports cover projects of public construction, including both technical and budgetal details, allowing sometimes the integral modern reconstruction of the buildings from the reports alone. The records of the Erechtheum and the Parthenon at Athens are well preserved, as are inventories of military expenditures, especially those of the Athenian navy. Knowledge of the ephebic system at Athens, a paramilitary youth organization, is in the main based on epigraphic material.

The laws of Gortyn

The only law code in the Greek epigraphic tradition is the laws of Gortyn in central Crete, inscribed on the slabs of a circular wall which, if completely preserved, would have been nearly 100 feet (30 metres) in diameter. The 12 columns of text, each on four layers of stone and some five feet (1.5 metres) high, are about 30 feet (nine metres) in sideways length and contain more than 600 lines of text, being the longest Greek epigraphic monument; parts of some columns of further text survive, the so-called Second Code. The probable date of this inscription is the first half of the 5th century BC. The code deals with such matters as disputed ownership of slaves, rape and adultery, rights of a wife upon divorce or death of husband, disposition of children born after divorce, inheritance, sale and mortgaging of property, ransom, children of mixed marriages, and adoption. While self-contained, it evidently does not represent the entirety of laws; curiously, it stresses those areas of civil law (inheritance, adoption) that are notably lacking in the Hittite Law Code. The uniqueness of this code in the Greek world points up the relative isolation and marginality of the Cretan tradition, with tendencies to codification more reminiscent of Anatolia and the Near East generally.

Ancient Rome. While partly overlapping Greek inscriptions in time and type, those of Rome nevertheless present distinct peculiarities. There is a high measure of standardization in kind and style, despite lingering local traditions in more remote areas. Extensive and excessive use was made of initials and abbreviations, to the point of serious impediments to comprehension; lists of such abbreviations are standard adjuncts to modern handbooks on Latin epigraphy. Stone and bronze were standard material, but there was more use made of bricks, tiles, and terra-cotta, and practices of stamping and signing such matter are of help in identification and dating.

Literary and epigraphic records of early republican Rome are scant and fragmentary. Latin was at the time still largely confined to Rome proper, with Oscan, Etruscan, and colonial Greek spoken and written in much of Italy. With the arrival of extended political power there was little early literacy to fall back on, and historiographic attempts at retrospection ended in epicized myth and legendry (*e.g.*, in Livy). The Greeks on the southern coastal fringe had little truck with the hinterlands of early times. The Etruscan impact on Rome is evident, but shortcomings in discovering epigraphic records of Etruscan city sites (as opposed to necropolises) and in understanding the Etruscan language, limit the historical data derivable from Etruscology. The potential for such illumination is seen from the discovery of gold tablets at Pyrgi in 1964 that contain a dedication in Etruscan and Phoenician by the Etruscan king of Caere, Thefarie Velianas, to the syncretized goddess Uni (Juno) Astarte. Datable to around 500 BC, the text shows Etruscans ruling in the outskirts of Rome, with enough Phoenician or Punic (Carthaginian) maritime presence to warrant symbiotic and syncretistic bilingualism. The vital historical import of such attestations, pieced together with later Greek and Roman historiographic data, is patently manifest.

Texts from republican Rome

No historically important epigraphic Latin text from republican Rome antedates the 2nd century BC. The marble Columna Rostrata—found in Rome in 1565 and now at the Palazzo dei Conservatori on the Capitoline Hill—records a naval victory of Duilius (consul in 260 BC) over the Carthaginians; but the inscription, replete with fake archaism, dates from a restoration effort in early imperial days. The *fasti consulares* and similar lists afford a summary sequence of consulates, magistratures, and triumphs. The one truly significant epigraphic historical text is the "Res gestae divi Augusti," an autobiographical record of Augustus' rule, which was exhibited in many places but is best known as the Monumentum Ancyranum, from the bilingual (Latin and Greek) version carved on the walls of the Temple of Rome and Augustus at Ankara (Turkey).

By the time the epigraphic record became abundant, Rome's domination was secure and the political documentation was one of imperial outflow and or local sycophancy. Treaties of republican Rome with foreign powers survive merely in the works of literary historians. Among "internal" documents from republican days are several epigraphic texts of significance: the Senatusconsultum de Bacchanalibus, on a bronze tablet found in 1640 in Brutium (the "toe" of Italy) and now in Vienna, is a consular edict on Senate authority, regulating Dionysiac outbursts in Italy in 186 BC; pieces of the laws Lex Acilia Repetundarum (123 BC) and Lex Agraria (111 BC) were found in the 16th century on opposite sides of what was once a large bronze tablet; the local laws of the town of Bantia (on the borderlands of Lucania and Apulia in southern Italy) are inscribed on a fragmentary bronze tablet found in 1790 (now in Naples), with a Latin-language text on one side and the longest known Oscan inscription on the other, both datable to the late 2nd century BC; parts of the Lex Cornelia de Viginti Quaestoribus (81 BC) are preserved on a large bronze tablet found at Rome; Julius Caesar's Lex Julia Municipalis of 45 BC was found near Heraclea in Lucania. On the whole, however, the transmission of Roman law, from the earliest fragments to the mature codifications, is nonepigraphic. In later times the flood of administrative decrees increases with the growth of centralized autocracy. Typically Roman epigraphic material of imperial date comprises further building inscriptions, military records, and honorific texts.

The oldest Turkic monuments

The Turkic peoples. The oldest monuments of Turkic languages—inscribed on stones, and datable to the early 8th century AD—were discovered in the late 19th century in southern Siberia around the Yenisey River and in northern Mongolia near the capital of Urga (modern Ulaanbaatar). Deciphered in 1893 by the Danish scholar Vilhelm Thomsen, they provide valuable insights into the history of Central Asia around the 7th century AD. These records of the Turk dynasty (Chinese T'u-chüeh) comprise especially texts found at Kosho-Tsaidam on the Orhon (Orkhon) Gol (river), including also Chinese text. These texts throw light on the nomadic culture of the tribal empire controlled by the Turk dynasty, including shamanism, calendar, customs, and social structure, with strong Chinese influence detectable in the latter. After the decline of the Turk people (*c.* 745), their successors, the Uighurs, perpetuated for a time the same kind of monumental dynastic epigraphy, the writing system of which is an offshoot of the Aramaic alphabet, presumably mediated by the Iranian-speaking Sogdians of Central Asia. Gradually, however, new scripts took over (especially the so-called Uighur alphabet, of Syriac origin, which was further transmitted to the Mongols and the Manchus) and inscriptional monuments gave way to manuscript records such as those found in Chinese Turkistan (Turfan) in the late 19th century (along with texts in Sanskrit, Sogdian, Tocharian, and other Indo-European idioms), attesting to a coexistence of Buddhist, Manichaean, and Nestorian Christian religious communities. The later Turkish peoples, including the Anatolian Seljuqs and Ottomans, had an Islāmic book-tradition, to which the inscriptional record is merely incidental.

Northern Europe. The advent of writing was slow north of the Alps; it came either from direct expansionary exportation by Greek coastal colonies and the Roman Empire, as in Gaul and the Iberian Peninsula, or indirect inspiration from the same quarter, as in writing in the Irish and British ogham alphabet and the Germanic runes.

Celt-Iberian inscriptions from Spain and Celtic ones from Gaul and Ireland are scarce, mostly brief, and notably devoid of usable historical information, apart from their mere monumental existence and linguistic and onomastic (pertaining to names) content. Occasional items such as the fragmentary Gaulish Calendar of Coligny afford insights into local cultural practices, apart from an overwhelming trend to romanization.

The runic alphabet—a Germanic alphabet, originally of 24 letters, also called futhark—and its offshoots (the Scandinavian, especially Danish, 16-letter variety from the 9th century AD; and Anglo-Saxon versions, from the 3rd to the 10th centuries AD, also called futhorc) are probably of "North Etruscan" or "Sub-Alpine" Italic inspiration, datable to around 200 BC. The "North Italic" letters of the Germanic text *harixasti teiva,* "to the god Harigast," on a helmet from Negau (southern Austria) are probably from that time of transmission. Runic inscriptions from the era of migrations, ranging from eastern France through Germany up to Denmark and eastward via Poland to Romania, are supplemented by the later, richer yield from England and Scandinavia. Native Anglo-Saxon runic epigraphy, mostly in Northumbria, Mercia, and Kent, petered out around the 10th century, whereas the Scandinavian tradition (including its enclaves on British soil) endured for several more centuries. Sweden has some 3,000 runic monuments; Norway and Denmark, perhaps 400 each; while Iceland has remarkably few, apparently in inverse proportion to the literary flowering in that colonial outpost. The Vikings left their runic calling cards in far-flung places, including those in the Greek port of Piraeus, on the Black Sea coast, in Varangian Russia, in Scotland, Ireland, and the Isle of Man, and the Orkneys, Hebrides, and Shetland islands; Greenland has its share, but alleged specimens from North America, notably the Kensington Stone from Minnesota, telling of the westward trek of an exploration party from Vinland, are invariably crude latter-day forgeries.

Runic monuments

The purposes of runic inscriptions were usually either dedicatory or commemorative, sometimes magic, and frequently sepulchral. The longest, that from Rök in Sweden (725 runes), seems to contain a catalog of epic deeds, possibly those of the Ostrogoth king Theoderic. The prime historical value of runic epigraphs is usually what and where they are, rather than what they depict or record.

Inscriptions as social and cultural records. In the preceding section, inscriptions were evaluated as sources for the presence and migrations of peoples, the existence and chronology of political states, their dynastic histories, foreign relations, internal governance, legal institutions, and official acts. In this section, epigraphy is surveyed for information about how past civilizations lived; their religious beliefs and practices; their business, financial, legal, and social relations; and what shape their aspirations assumed in terms of verbal creativity. The subdivision of civilizations surveyed differs somewhat arbitrarily from the earlier section by the omission of certain areas and the inclusion of Crete and Mycenaean Greece. In fact, Indic and Chinese epigraphic matter discussed above could just as well have fitted the "religious" slot, but its royal character and chronological importance for official history dictated otherwise. Conversely, the Cretan and Mycenaean tablets are purely economic inventories, but they might possibly have been included with history above for the very important historical fact that they prove the ruling presence of Greeks at Knossos during the 2nd millennium BC. The varying degree of importance of epigraphic material in various cultures persists: in Mesopotamia and the ancient Near East its dominance was nearly total; in Egypt it combined with the papyrological dimension; in Crete it was merely a flash in a prehistoric darkness; while in ancient Greece and Rome, it was a supplementary concomitant of the nonepigraphic literary tradition.

Ancient Mesopotamia. Ample specimens of Akkadian-language clay-tablet epistolography have been found at several sites, notably Tell el-Amarna in Egypt and Tell al-Ḥariri on the middle Euphrates (the ancient Mari of *c.* 1700 BC). The Amarna letters, about 400 of them, were composed in corrupt Akkadian by Canaanite scribes in Syria and Palestine and were largely official in character. The Mari letters, some 5,000 in number, are more illustrative of normal day-to-day written communication in a Mesopotamian milieu proper.

The Amarna and Mari letters

Another aspect of everyday life in ancient Mesopotamia is amply illustrated by thousands of clay tablets of a practical legal nature, as distinct from the formal laws. These archivally preserved records from various periods use Sumerian, Old Akkadian, Babylonian, and Assyrian alike. They detail law suits, court decisions, marriage contracts, divorce proceedings and settlements, sale adoptions (fictitious acts circumventing prohibitions against land sales outside the family), loan agreements, tax receipts, and much else. Commercial inventories, such as those of the Old Assyrian merchant colony at *Karum Kanes* in central Asia Minor (20th century BC), complete the picture.

Due to the religious sanction of law, legal records were often stockpiled in temple archives. These latter are also the source of more directly cultic texts, such as descriptions of rituals, which come under such headings as "Temple Program for the New Year's Festivals at Babylon," "Ritual to be Followed by the Kalû (priest) when Covering the Temple Kettle-Drum," "Ritual for the Repair of a Temple," and "Program of the Pageant of the Statue of the God Anu at Uruk." Prayers, lamentations, and hymns in both Sumerian and Akkadian are extant, addressed to deities such as the goddess Ishtar, the moon god Sin, the sun god Shamash, the great triad Anu, Enlil, and Ea, and the Babylonian patron god Marduk. The Sumerian "Lament for the Destruction of Ur" bemoans the city's fall to Elamites and Subarians. Often the king himself is the spokesman in the text. Wisdom literature, such as proverbs and fables (*e.g.,* "Dispute between the Date Palm and the Tamarisk"), poetic meditations, oracles, divination records, omens, and prophecies are further examples of Mesopotamian genres that only epigraphy has preserved.

Sumerian and Akkadian narrative literature is likewise of wholly inscriptional transmission. It contains man's earliest preserved literary creations in the Sumerian sequence, especially the texts from tablets found at Nippur. These include the "Paradise myth" of the god Enki and the goddess Ninhursag in the pure, clean, and bright land of Dilmun; the story of Dumuzi and Enkimdu (the petulant shepherd god versus the peace-loving farmer god, inversely reminiscent of the Cain–Abel antagonism in Genesis but not culminating in murder); "The Deluge" with its Noah-hero Ziusudra; "Inanna's Descent to the Nether World," which prefigures the later Akkadian "Ishtar's Descent"; and the lays of Gilgamesh, which show the Sumerian traditions that were later partly organized and transformed into the Akkadian epic. The latter include "Gilgamesh and Agga of Kish," a story of confrontation between early Sumerian city-states; "Gilgamesh and the Land of the Living"; and "The Death of Gilgamesh," with its haunting parallelistic refrain "he lies, he rises not." The Akkadian *Epic of Gilgamesh* is divided into 12 tablets, the longest of which is more than 300 lines; this "Flood Tablet" (the 11th) is virtually intact and comes, like almost all Assyrian-language Gilgamesh texts, from the library of Ashurbanipal at Nineveh (7th century BC). From the 2nd millennium there are fragments of a Hittite version from Boğazköy, as well as minor traces of a Hurrian translation. Old Babylonian correspondences to tablets 1–3 and 10 are found on a tablet from Sippar (*c.* 1800 BC). The 12th tablet is a literal translation from Sumerian, whereas the rest amounts to a self-contained Akkadian epic original, based on Sumerian motifs but with a thrust of its own. The most complete reconstruction involves a combination of Assyrian, Old Babylonian, and Hittite versions.

The Gilgamesh epic

The other famous Mesopotamian epic, *Enuma elish,* "When on high," details the story of cosmic creation and of how Marduk became the great god of Babylon; it had

more immediate cultic attachments because its recitation formed part of the New Year festival.

Further Akkadian literary creation is attested in the epic of Atrahasis, a tale of mankind's punishment through pestilence and flood, preserved in fragmentary Old Babylonian and Assyrian versions. The story of Adapa, found in parts in the Tell el-Amarna archives and the library of Ashurbanipal, is similar to Gilgamesh's quest for immortality. The myth of Zu deals with the theft of the tables of fate and the usurpation of almightiness by the bird god Zu. The legend of Etana, a namesake of the shepherd-king who ascended to heaven in the mythical postdiluvian Sumerian dynasty of Kish, recounts in its Old Babylonian and Assyrian recensions the heavenly flight of Etana on the wings of an eagle in order to acquire the magic birth plant that would cure his childlessness. Death-oriented themes appear in the tale of Ishtar's descent, in the story of Nergal and Ereshkigal, and in various netherworld texts associated with the Tammuz myth and liturgy.

The accidental preservation of Egyptian texts

Ancient Egypt. The multitude of incidental day-to-day written matter on potsherds and papyrus, preserved by climatological quirks in Egypt, affords insights into the normal living patterns of ordinary people during the pharaonic, the Hellenistic (Ptolemaic), and the Roman periods. More formal records of similar provenance may be described as legal in the widest sense, comprising such objects as land deeds, cadastral inventories (tax surveys), wills, adoption decrees, and trial transcripts. Thus many technicalities of landholding in the Nile Delta during certain ancient periods survive, unwarranted by any normal expectancies of epigraphic preservation elsewhere; only the happenstance of the Creto-Mycenaean records (see below *Crete and Mycenaean Greece*) has accidentally transmitted something analogous.

Ritualistic texts are equally abundant, in both monumental and papyrological transmission. Notable among offertory liturgy was the mortuary service to the dead, acted out as a kind of ritual make-believe in which the offering was referred to as the eye of Horus (the sacrificer) being given up on behalf of his father, Osiris (the departed). Ritual drama in general is amply attested in Egyptian religious practice; day-to-day ceremonial in the cult of gods or god-kings (pharaohs) was recorded in the minutest detail. Magic texts include incantations and charms against such dangerous creatures as snakes, scorpions, crocodiles, and lions. Even the buried dead were deemed in need of such protection, and hence the preserved samples are largely of mortuary provenance. An important category of epigraphs is constituted by curses, imprecations, and threats of divine vengeance against enemies, evildoers, and tomb violators in particular. Dream interpretation and oracular inquiry are represented on both monuments and papyri.

Of a more secular nature and on the verge of true literature are moral and didactic writings, particularly during the early Middle Kingdom (began 1938 BC), when a profound social and spiritual crisis seems to have gripped Egypt. Of such kind are "The Admonitions of Ipuwer" (a denunciation of current sin and evil in Hebrew "prophetic" manner), the "Colloquy of a Prospective Suicide with His Own Soul," and especially copious compendia of conventional wisdom, frequently formulated as instructions from a royal or noble father to his son. Some parts offer remarkable parallels to the Hebrew book of Proverbs.

On the same borderline to real literature stand mythical tales, including in particular the myths of creation by the god Atum of Heliopolis and the Memphite patron deity Ptah, whose eminence is reminiscent of the cultic fortunes of Marduk. The Memphite cosmogony, of 1st dynasty provenance, is an abstract one, with Ptah's creation conceived with his "heart" (mind) and "tongue" (word) rather than the usual physical detail. Other mythic tales concern the perils from the netherworld demon Apepi faced by the sun god Re during his nightly journey in his solar boat; the rivalries between Horus and Seth for the succession of Osiris (a patent allegory of the tensions between Lower and Upper Egypt); mortuary texts particularly concerned with the Osiris myth and the prospect of judgment (*e.g.*, the "Protestations of Innocence"), attested both monumentally and on papyri (the Book of the Dead); and pieces designed to bolster the supremacy of specific sites and deities, particularly Amon-Re at Thebes.

Out of the latter tendency emerged hymnic poetry during the 18th and 19th dynasties, triggered by an imperial trend toward unicentrism in government and universal syncretism in religion. The great hymns to Amon-Re were matched during Ikhnaton's "Amarna revolution" by his great Aton-hymn, uncannily reminiscent of the 104th Psalm of the Old Testament. Victory hymns and prayers are further examples of poetic genre, as are purely secular pieces, including several collections of love songs from *c.* 1300 to 1100 BC, songs of farm workers and herdsmen, and feast songs for entertainment.

In nonreligious literary prose there is preserved the story of Sinuhe, an account of the life of a Middle Kingdom official, his prosperous peregrination in Asia but continued yearning for the homeland, relieved at last by a chance to join the pharaonic court. Papyri and ostraca tell this story in a long succession of attestations through most of the 2nd millennium BC. Of later date (*c.* 1100 BC) is the "picaresque" narrative of the hieratic-script "Moscow papyrus," detailing the ill-financed mission of a certain Wen-Amon from Karnak to Byblos to buy timber for a ceremonial barge of Amon.

The Boğazköy archives

The Hittite Empire. The Boğazköy archives are the unique central storehouse of Hittite records for the duration of that empire (only minor additions have been found elsewhere, such as letters at Tell el-Amarna and Alalakh). The past of other cultures was known from external and nonepigraphic sources before the exploration and excavational discoveries of inscriptions; with the Hittites, however, the record is in one place and *in toto.*

Much may be learned of everyday life and social relations from legal texts, including the laws themselves. Such matters as family relations, property rights, land tenure, and commodity prices are dealt with in great and sometimes pedantic or amusing detail. More can be learned from gift deeds of land, cadastral lists, and trial records, the latter containing largely depositions of respondents and witnesses in civil suits. Unique in their kind are treatises on the training of racehorses. Astrological omen texts, divination records (hepatoscopy, auspicy, lottery oracles), and dream interpretations abound, patterned largely on Mesopotamian models; rituals of all kinds take up a disproportionate amount of space. Temple foundations and procedures, festival ceremonies, epidemics, impotence, family quarrels, countermagic, royal welfare, birth, insomnia, and the manipulation of deities are among the topics of Hittite ritual; in a class by itself is a funeral ritual for royalty, reminiscent of Homeric practices. Not only the Hittite language but Hattic, Hurrian, Luwian, and Palaic are used in rituals. Another large category is made up of hymns and prayers, most notably those of King Mursilis II, in particular his "plague prayers," in which he confesses sins and begs the storm god for mercy. Other notable examples are the prayer of Muwatallis to the sun god, the vows of Queen Puduhepa to the sun goddess, and the prayer of Kantuzilis with its intimations of mortality. Native Anatolian mythical texts, such as those of the slaying of the dragon Illuyankas and of the disappearance of the god Telipinus, were part of ritual recitation. Other mythical matter points up the crossroads character of Hittite civilization, being made up of Akkadian, Hurrian, and Canaanite themes. Most of the Hittite epigraphic corpus has some connection with religion.

The city-state of Ugarit

Other countries of the ancient Middle East. Away from the big power centres some quite important sites remain to be identified, such as the Mitannian capital of Wassukkani. Others (*e.g.*, the fortress-town of Carchemish on the upper Euphrates, for some time a Hittite dependency) have yielded notable data, especially royal inscriptions in Hittite hieroglyphs. In particular the French excavations at Ras Shamra on the Syrian coast since 1929 have uncovered the inscriptional and other remains of the small but strategic city-state of Ugarit, which flourished in the 15th–13th centuries BC. Its own vowelless cuneiform alphabet, a radical departure from the Mesopotamian syllabary prototype, denotes an archaic Semitic dialect closely akin to Canaanite. The written documentation of this crossroads

community, in a variety of scripts and languages, provides information on the history of the area where Hittite and Egyptian power politics collided. But its most significant epigraphic products have been poetic epics of mythical and heroic scope, which go far to elucidate religious traditions otherwise known mainly from Old Testament bias. The great cycle of Baal and Anath brings to life the pantheon to which belonged also El, Asherah (the Astarte of the Phoenicians), Kothar ("Deft," the craftsman god), and Yamm (the sea god). The tale of Aqhat is also on the borderline of myth, while that of King Keret (or Kret) takes place in a saga world somewhat reminiscent of the Greek Homeric tradition.

Scattered dedicatory inscriptions and papyrus texts shed some light on the lives and practices of the Aramaean-speaking populations during the 1st millennium BC. From Palestine there are the Hebrew ostraca of Samaria, datable to the reign of Jeroboam II of Israel (8th century BC), which record names, families, and administrative and religious practices. Of equal significance are the ostraca of Lachish in southern Palestine, which probably immediately preceded the Chaldean onslaught of 589 BC. Phoenician texts are scattered around the Mediterranean, and bear witness to an extensive and protracted maritime supremacy.

The Old Persian Achaemenid inscriptions (see above *Ancient Iran*) are of some value in assessing the state religion of the time, which seems to have been a rather anemic form of official Zoroastrianism. Later monuments from the wider Iranian area help map its complicated religious history, such as the great inscription of the Kuṣāṇa king Kaniṣka, found at Surkh-Kotal in Afghanistan in 1957 and attesting to the Iranian language and Mithraic cults of ancient Bactria in the 2nd century AD. The Sāsānian religious tradition of the 3rd–7th centuries, with its rigidified formal Zoroastrianism, is mainly of nonepigraphic attestation.

Crete and Mycenaean Greece. The decipherment of the latest and most copiously attested of the Minoan linear scripts, the so-called Linear B, by British cryptologist Michael Ventris in 1953, is a major example of the dramatic impact epigraphic discovery can have on the most varied antiquarian disciplines. It supplied incontrovertible proof that the Mycenaeans on the Greek mainland during the 2nd millennium BC were Greek in language and likewise that Knossos in Crete was a Greek-speaking stronghold at the time of its final destruction (*c.* 1400 BC). The records from Knossos, Pylos in Messenia, and Mycenae in the Argolid are exclusively perishable inventories on clay tablets, kept in royal palaces or emporia and apparently meant to be discarded or reused annually; thus they record data from the very last year of any given establishment, just before its final destruction in the fire that also accidentally ensured their preservation by baking the clay.

The Linear B script

The contents are brief, concise, practical, and quantifying in character. Nevertheless, they reveal an important amount of detail about a civilization that is otherwise accessible only via the testimony of archaeology, iconography, and the poetic memory of early classical Greece across several "dark" centuries. Apart from the philologically invaluable recording of preclassical linguistic forms and proper names, there is the cryptologic evidence of the syllabic writing system, which may yet help bring about a cogent decipherment and linguistic identification of the earlier forms of Cretan writing.

Something of bureaucratic accounting methods can be learned from the system of the tablets. Contents yield lists of personnel; of livestock and agricultural produce; of textiles, vessels, and furniture; of metals and military matériel (swords, corselets, chariots, etc.); as well as records of landholding and of temporal and cultic tribute. The cadastral inventories are significant for the light they shed on the system of land distribution as compared with classical Greece and especially with contemporary practices in adjacent culture areas, such as those recorded in the Hittite Law Code. The religious tributes are revelatory in proving what bold scholarship had merely surmised—that the bulk of the classical Greek deities existed in name and kind already in Mycenaean days. Something of the social order and the system of government can also be incidentally learned.

Classical Greece. Inscriptions in their variety and profusion are an important means of gauging everyday life in classical Greece, especially such more formal aspects as were deserving of recording. This qualification, however, does not prevent some obscene pederastic rock-carvings on the island of Thera from being the earliest epigraphic texts from the classical period (*c.* 8th century BC), in a form of the alphabet still strongly marked by its Phoenician source.

Earliest classical epigraphic texts

In a formal vein there is ample documentation of a business, financial, and legal type, such as regulations for building projects, contracts for medical services (*e.g.*, the Cypriot bronze tablet of Idalium, in a peculiar syllabaric script reminiscent of the Cretan and Cypro-Minoan types, used to write both Greek and the uninterpreted Eteocypriot language), collection agreements for defaulted notes, manumission decrees, records of bank deposits (frequently in a temple), and reports of land commissions (especially the Heraclean Tables from southern Italy).

Even in these formal documents it is impossible to keep out reference to religious establishments. Religion was, in fact, big business in classical Greece, and overlap in the records is frequent. Temples were often of the order of corporations, with extensive holdings of real estate and banking operations. The epigraphs of their transactions are coupled with minute records of dedicatory offerings, priest lists, administrative regulations, and ritualistic and liturgical instructions. So much of social life was centred around institutionalized religion that most of its documentation may be included under the latter heading. Dedicatory texts in the widest sense are common, often in verse, such as that by Nikandre of Naxos on a 7th-century-BC statue of Artemis at Delos, a kind of propitiatory offering to the maiden goddess on the occasion of marriage. Similar hexametric or elegiac texts are found as epitaphs, especially for those who perished in war or at sea.

Tomb inscriptions generally are the most numerous of all but tend to be laconic in style and lacking even in purely statistical data. Those from ancillary areas of the Greek-speaking world, such as the non-Greek ones from Lycia, Lydia, and Phrygia in Asia Minor, are sometimes more revealing to the extent that they are interpreted. Thus the Lycian stela of Xanthus, a tomb monument from the 5th century BC, runs to many hundreds of words, including a dozen lines of Greek. The peculiar Lycian system of matrilinear descent is clearly evident in the texts. The Lydian ones include a marble stela from the necropolis of Sardis with a Lydian-Aramaic bilingual epitaph that typically calls down the curse of Artemis of Ephesus on potential violators. Such imprecations are also standard in Phrygian sepulchral epigraphs. Some kinds of death-related inscriptions were intentionally buried, such as the curses that consigned the object to the infernal avengers. Another type is seen in the Orphic tablets—texts on thin gold found interred with the remains of devotees of the Orphic salvation religion near Sybaris in southern Italy, near Rome, and at Eleutherna in Crete. They are apparently meant as instructions to the deceased—exhortations to the soul on its progress in the beyond, enjoining the shunning of the spring of Lethe ("Oblivion") and the drinking from that of Mnemosyne ("Memory"), thus securing the end of transmigration of the soul and the attainment of perpetual higher consciousness.

Tomb monuments

The Orphic texts provide the popular parallel to the Platonic myth of Er in the 10th book of *The Republic*, where Plato somewhat recast Lethe and Mnemosyne into Ameles ("Unmindfulness") and Anamnesis ("Recollection"). Thus, epigraphy not merely supplements the literary record of Greek civilization but also complements it in important aspects. If papyrology is included, important additions to the preserved portion of classical Greek literature have come from Egypt, especially in the papyri from Oxyrhynchus; these include such otherwise lost treatises as Aristotle's *Constitution of Athens* and entire dramatic works (*e.g.*, the *Dyscolus* of Menander).

Ancient Rome. The frequency of Roman inscriptions increased dramatically in direct proportion to the rise of

Roman power; but that same rise brought centralization, stereotyping, and a certain sterility of the more formal parts of the epigraphic record. Much of the bulk is official, unidirectional, and from the top outward. The Greek-speaking parts of the empire continued in many of their ancient ways, and other annexed regions (such as Spain or Gaul) developed their own locally coloured practices. The spread of a variety of exotic religious cults all over the empire, among them Mithraism and Christianity, added a kind of epigraphic underground initially devoid of official sanction and largely unmatched by alternative avenues of preserved written transmission. Popular epigraphy, including such matter as graffiti at Pompeii and other Vulgar Latin inscriptions, provides further counterpoise to the official stereotypes.

From early republican days the Roman written record is very spotty. The earliest text of any length, the Forum inscription from the early 5th century BC, seems to refer to an augural rite (referred to by Cicero and Festus as the *juge auspicium*) which enjoins the immediate unyoking of beasts of burden that produce excrement. A 4th-century vase (the "Quirinal vase") is apparently a wedding present with an inscribed guarantee by the bride's guardian to use his good offices to further marital concord. Later texts of religious import include the "Song of the Arval Brethren" on a marble tablet found in the Vatican in 1778 (a chant largely reduced to gibberish by age-old ritualistic repetition), records of various other cultic "colleges" (*augures; fetiales,* or priestly diplomatic representatives; *salii,* or priests of Mars and Quirinus et al.), and outside of the Latin-speech area especially the seven Umbrian-language bronze tables found at Iguvium (modern Gubbio) in 1444, recording in more than 4,000 words the ritualistic details of a brotherhood (the Fratres Atiedii) that flourished in republican days. Official religion is further seen in many dedications, such as one found on Caelian Hill in Rome in the 18th century, which records the offering of a temple and statue to Hercules by Lucius Mummius in 146 BC in fulfillment of a vow he had made during his conquest of Greece and sack of Corinth. "Unofficial" cults, especially those of Mithra and Jupiter Dolichenus, can be traced and mapped throughout the empire chiefly with the help of dedicatory epigraphs. Sepulchral inscriptions are recorded in overwhelming numbers, not only in Latin but notably also in Etruscan; among the latter is the mummy wrap, a lengthy funeral liturgy inscribed on an apparently Etruscan mummy entombed in Egypt (now in Zagreb). The earliest notable Roman examples are the epitaphs in Saturnian verse on the sarcophagi of the Scipios from the 3rd century BC.

There are many texts of deep sentiment and poetic feeling. Yet the bulk itself is the main source of information because it affords what may be called demographic statistics, including population distribution, occupations, public health, and longevity. The frequent inclusion of the *cursus honorum,* or career summary of the departed, has similar informational value. Tombs were usually placed under the divine tutelage of the powers of the beyond. Imprecations against violation were common, as were buried curse tablets in Latin and in Oscan (the *defixiones*). Many of the unofficial inscriptions exhibit substandard degrees of literacy and colloquial features, being thus an important adjunct to the study of the totality of classical Latin and its subsequent developments.

The use of inscriptions. *The dating of historical events.* Inscriptions are important specimens for chronology because they are often physical objects contemporary in execution with their contents. The dating of the inscription itself frequently yields a trustworthy chronology of its message: a victory stela records something freshly deserving of celebration; an epitaph implies a recent death. Exceptions do exist, which record more or less remote events at a conscious historical remove; archival specimens, for example, and secondhand copies generally lack the contemporaneity of other inscriptions. On the whole, however, external dating is crucial and may be achieved in several ways. Excavated monuments can be chronologized by their archeological context, including stratigraphic analysis and radiocarbon dating of any adjacent remains of organic matter. The shape of the monument may permit stylistic and iconographic determination. The type and variety of script used, and especially the style of writing, often allow paleographic dating. Thus, the relative age of Hittite texts can be determined by spotting the typical "Old Hittite ductus" of the more ancient period, and the various "scribal hands" of the Linear B tablets have been differentiated with extreme subtlety. Sometimes a radical reform, such as the official adoption by Athens of the Ionic alphabet in 403 BC (replacing the local Attic variety), provides a chronological watershed. Internal evidence of the inscription may yield its own kind of dating, either by synchronism with otherwise known facts or events or in true calendaric fashion. The year is frequently indicated by a king's reign or the tenure of a magistrate. Such stable counting of time as, for example, from Rome's legendary founding in 753 BC, or from creation (5509 BC, according to Christian dating) is rare in inscriptions; it became more of a historian's device during the classical and postclassical periods. In smaller communities, however, especially in Asia Minor, analogous local departures were used (legendary or historical foundation dates or other epoch-making events), with confusing results for latter-day chronologists.

Contemporaneity of epigraphy

The use of inscriptions for the dating of historical events is most pervasive when the historical tradition itself is "timeless," as in ancient India; the entire Indian chronology comes to be anchored around the Aśokan inscriptions. Inscriptions also permit a check on the veracity of ancient historians such as Herodotus (dubbed both "father of history" and "father of lies"), as in the case of the Bisitūn inscription of Darius. Equally dramatically, the Linear B tablets prove at one stroke that the Greeks were ensconced at Knossos in the 2nd millennium BC and that the bulk of the "Olympian" religion was already theirs at the time. Most significant of all, nothing would be known of the great Hittite Empire during the 2nd millennium BC, were it not for the discovery of its inscriptional archives.

The decipherment of ancient languages. Inscriptions as written records are usable only in proportion to their intelligibility. Important epigraphic corpora remain virtually undeciphered; *e.g.,* the "Indus script" from Mohenjo-daro and Harappā (3rd millennium BC), the Carian texts from Asia Minor and by Carian mercenaries in Egypt (1st millennium BC), and the pictographic Mayan "hieroglyphs" from Central America (*c.* AD 500–1500). Sometimes the writing system is intelligible, as in the case of Etruscan, but understanding remains deficient because the language is otherwise unknown and bilingual keys are lacking or inadequate. Chances for success are best if there are sufficiently extensive bilingual or multilingual copies, of which at least one language is previously understood. Such presence made possible the decipherment of ancient Egyptian (the Rosetta Stone of 196 BC, with hieroglyphic, demotic, and Greek versions). Once the affinity of an underlying language to known idioms is established (*e.g.,* Old Egyptian to Coptic, Old Persian to Avestan and Sanskrit, Akkadian to Hebrew), interpretation can proceed apace. The recovery of Hittite was not a true decipherment because the script was a relatively common variety of syllabic cuneiform. The interpretation was helped by the nature of the writing on the one hand (including intelligible ideograms, while an alphabet yields no such clues), and by the presence of Akkadian-Hittite bilinguals on the other; the soon-recognized Indo-European affinities of the Hittite language afforded further help. The Hittite hieroglyphs were partly deciphered by painstaking internal analysis based on the correct assumption of an underlying dialect akin to Hittite; a bilingual with Phoenician text brought much welcome confirmation. The decipherment of Linear B was a sheer triumph of methodical cryptology, again based on the correct hunch that the hidden language was Greek.

Undeciphered epigraphic corpora

In sum, the decipherment of ancient scripts and the recovery of lost languages are practically identical with the interpretation of previously unintelligible ancient inscriptions, because texts involved are almost exclusively epigraphic. Even when the language is otherwise preserved (as with classical Greek and Latin), inscriptions yield es-

sential additional data for its history, dialects, and social diversification.

History of epigraphy. *Greek and Latin inscriptions.* Inscriptions have commonly elicited the curiosity of posterity, and such ancient Greek historians as Thucydides and Polybius already made scholarly use of them. Sporadic systematic interest in Greek and Latin inscriptions is attested in later ages; *e.g.*, Cola di Rienzo in the 14th century made a collection, and Cyriacus of Ancona (Ciriaco de' Pizzicolli) in the 15th century was a renowned recorder of ancient written monuments on his mercantile travels to Greece, Anatolia, and Egypt. Cyriacus' material formed the nucleus of various compilations in the succeeding centuries, normally on a geographic basis. A rival typological method of publication was launched by Martin Smetius in Leiden in the late 16th century and was followed in the early 17th by a monumental collection of Janus Gruter and Joseph Justus Scaliger. After copious additions of material during the 18th and early 19th centuries, the *Corpus Inscriptionum Graecarum* was launched by August Böckh in 1815 under the aegis of the Berlin Academy and was completed in four volumes with index (1828–77). The material had by then again outrun the publication, and it was resolved in 1868 to re-edit completely all Attic inscriptions. Ulrich von Wilamowitz-Moellendorff in 1902 took charge of the *Inscriptiones Graecae* (1873–), which continued where the *Corpus Inscriptionum Graecarum* left off and included the *Corpus Inscriptionum Atticaru,* as well as all Greek inscriptions from European Greece (including Magna Graecia in Italy) and Cyprus. Those of Anatolia were left to the *Tituli Asiae Minoris* of the Vienna Academy, which began with the Lycian-language inscriptions from Lycia in 1901 and continued with the Greek and Latin ones from Lycia in 1920–44. The rest of Greek Anatolia was combed somewhat by the multivolume American series, *Monumenta Asiae Minoris Antiqua* (since 1928). *Inscriptiones Graecae,* framed in 14 volumes, has turned partly into a kind of overall umbrella for diverse coverage; volumes 6, 8, 10, much of 11, parts of 12, and 13 were never completed, being preempted by such other large publications as *Inschriften von Olympia, Fouilles de Delphes,* V. Latyshev's *Inscriptiones Antiquae Orae Septentrionalis Ponti Euxini Graecae et Latinae* and G. Mikhaylov's *Inscriptiones Graecae in Bulgaria Repertae, Inscriptions de Délos,* and M. Guarducci's *Inscriptiones Creticae.* The continuous influx of new material is inventoried and presented in such annual outlets as *Supplementum Epigraphicum Graecum* (since 1923) and "Bulletin épigraphique" in the *Revue des études grecques.* Large recent undertakings either supplement or complement the *Inscriptiones Graecae,* such as *The Athenian Tribute Lists* (from the American excavations in the Agora) and *Inscriptions grecques et latines de la Syrie* (six volumes, 1929–67). There are, in addition, numerous more localized or specialized collections.

Epigraphic collections

The *Corpus Inscriptionum Latinarum,* founded by Theodore Mommsen in Berlin in 1862, comprises 16 volumes fortified with copious supplements; it covers the classical Latin world and is notable for its homogeneity and systematization. Current finds and studies are presented in *L'année épigraphique* (since 1888). Etruscan inscriptions are less perfectly published in the antiquated *Corpus Inscriptionum Etruscarum;* many important later finds are included in transcription in M. Pallottino's *Testimonia Linguae Etruscae* and in M. Fowler and R.G. Wolfe, *Materials for the Study of the Etruscan Language* (1955; a computerized corpus). Of further relevance to the Roman world are collections of the type of E. Diehl, *Inscriptiones Latinae Christianae Veteres;* M.J. Vermaseren, *Corpus Inscriptionum et Monumentorum Religionis Mithriacae;* and *Inscriptiones Italiae.*

Other inscriptions. R.A.S. Macalister's *Corpus Inscriptionum Insularum Celticarum* (1945–49) gathers the ogham and other early texts from Ireland and elsewhere. The runic inscriptions are inventoried in a variety of compilations. The *Corpus Inscriptionum Semiticarum,* in Paris (since 1881), covers in separate volumes Phoenician, Aramaic, and other speech areas. Urartean texts were collected by C.F. Lehmann-Haupt in *Corpus Inscriptionum Chaldicarum* (1928–35); the earlier found Hittite hieroglyphic texts, by L. Messerschmidt in the antiquated *Corpus Inscriptionum Hettiticarum.* The *Corpus Inscriptionum Iranicarum,* intended to gather the epigraphs of Persia proper (Achaemenid, Seleucid, Parthian, Sāsānid) and of eastern Iran and Central Asia, began in London in 1955. The *Corpus Inscriptionum Indicarum* has published four volumes since the 1870s, comprising the Aśoka, Indo-Scythian, Gupta, and Kalacuri-Cedi periods, supplemented by the series *Epigraphia Indica* and *South Indian Inscriptions.*

No equally closed corpora exist for cuneiform and Egyptian documents. The Hittite texts come closest to such organization, issued mainly in two great ongoing series, *Keilschrifttexte aus Boghazköi* (German Oriental Society) and *Keilschrifturkunden aus Boghazköi* (Berlin Academy). But mostly the publication is haphazard—as part of excavation records (*Keilschrifttexte aus Assur*), depending on storage sites (*Cuneiform Texts* in the British Museum), or on the basis of genres (F. Thureau-Dangin, *Rituels accadiens*). The same pattern holds true of the Egyptian records.

Much epigraphic material is published in excavation reports and in many archeological and antiquarian periodicals; *e.g., Hesperia, Journal of Hellenic Studies, Bulletin de correspondance hellénique, Archeologia classica, Studi Etruschi,* and *Syria.* Specifically epigraphic serial publications include *Epigraphica* (Milan, since 1939), *Kadmos* (Berlin, since 1962), *Zeitschrift für Papyrologie und Epigraphik* (Bonn, since 1967), and *Chiron* (Munich, since 1971). (J.Pl.)

GENEALOGY

The word genealogy comes from two Greek words—one meaning "race" or "family" and the other "theory" or "science." Thus is derived "to trace ancestry," the science of studying family history. The term pedigree, used to describe a genealogy as set forth in chart or other written form, comes from the Latin *pes* ("foot") and *grus* ("crane") and is derived from a sign resembling a crane's foot, used to indicate lines of descent in early west European genealogies. Chart pedigrees, familiar to most people from school history books, include arrow shapes, parallel lines, a crinkled line denoting illegitimacy, and the sign = denoting marriage. Genealogy is a universal phenomenon and, in forms varying from the rudimentary to the comparatively complex, is found in all nations and periods. In this section the history of genealogy will be outlined, followed by an account of the work of modern genealogists, professional and amateur, and as organized in associations.

History of genealogical study. The history of genealogy can be divided most easily into three stages. The first is that of oral tradition; the second, that in which certain pedigrees were committed to writing. The third stage comprises the period from approximately 1500 in western Europe and later in the English-speaking world, during which the whole basis of genealogy widened to such an extent that it is now possible for the majority of people in western Europe to trace their ancestry.

Oral tradition and biblical sources. In the early days of civilization, before written records were made, oral traditions were necessarily important. Without the art of writing, reliance must be placed on memory, aided possibly by mnemonic systems like that of knot arrangements used by the pre-Hispanic Peruvians, or beads employed by the Maori of New Zealand. The ancient Scottish sennachy, or royal bard, could recite the pedigree of the old Scots kings at the latter's inauguration, and the nobles of Peru, who boasted a common descent with the sovereign, were able to preserve their pedigrees despite the complexity resulting from the practice of polygamy. Oral transmission of genealogical information is almost always as a list of names—the lineages of the ancient Irish kings, for example. Events of outstanding importance are occasionally incorporated in such lists.

Numerous Oriental genealogies appear in the Old Testament. A cursory examination of these will reveal that they belong to the first and second stages in the history

Oriental genealogies

of genealogy, as described above. The systematic keeping of genealogical records, as in Europe since 1500, did not occur until very recently in Asia and Africa.

In southern India the ruling house of the maharajas of Travancore claimed to trace its descent, direct and unbroken, from the old Cēra kings of southern India (referred to as independent sovereigns in one of the edicts of Aśoka, the great Mauryan emperor of the 3rd century BC). A claim that inscriptions of the rulers of Travancore have been found from the 9th century AD comes from a statement issued by the secretariat of the maharaja of Travancore. Its reliability may be judged along with the genealogies of princes in northern India shown in Lt. Col. James Tod's monumental work, *Annals and Antiquities of Rajasthan* (1829, republished 1950). Referring to the lineages of Indian princes as being known since the early centuries BC, Tod wrote, "If, after all, these are fabricated genealogies of the ancient families of India, the fabrication is of ancient date, and they are all they know themselves upon the subject." The very long Oriental genealogies begin as oral pedigrees and were later written down, but they concern only princes or great persons.

In Africa the one instance of a claim to very long descent, that of the Emperor of Ethiopia, bears a similarity to Tod's Rājput genealogies. The Emperor is said to descend from the marriage of King Solomon with the Queen of Sheba. The tradition was written down more than 15 centuries ago; it is therefore older than the history of most European monarchies, but it cannot, of course, be substantiated by documentary proof.

Under European influence the greater Oriental countries have adopted the practice of keeping systematic records for all citizens. In China, with its ancient system of ancestor worship, long, drawn-out pedigrees, including claims to descent from Confucius, are not unknown. The establishment of the Chinese Republic in 1911 brought with it registration of vital statistics.

In modern Japan, as might be expected from its thoroughgoing Westernization, the registration of vital statistics is regulated by law. The Family Registration Law of 1947, and later enactments, require a comprehensive registration of a Japanese national from his birth to his death. Such information, however, is kept in local registration offices, and there is no system in Japan for gathering together, recording, and preserving the information in one central place (although of course the results of statistics, such as the number of births, is known to the central authority). Such an exact system of registration covers only the era of modern Japan. The present-day pedigree of the Japanese emperors has a divine origin; it is mainly a string of names, easily recited and memorized, mixed with semifabulous legends and first written down in the early centuries of the Christian Era. It is concerned only with exalted persons, royal or noble.

Biblical genealogies

In the Old Testament there are many genealogies, the object of which is to show descent from Adam, Noah, and Abraham. By the time these genealogies had become part of the Jewish scriptures, the concept of racial purity had reinforced the keeping of family records. Genealogies of Jesus Christ in the New Testament aim at showing his descent from David, the one in St. Luke's Gospel going as far back as Adam, "who was the son of God." The idea of divine origin was reflected everywhere in a wildly polytheistic form among the Gentiles. Almost without exception, the heroes whose genealogies were recited by the bards had their paternity ascribed to the gods, or to persons such as Romulus who were regarded as having become divine. Greek fables abound in stories of great men begotten by gods and mortals.

In Roman genealogies heroes were always descended from gods. Julius Caesar, for example, was supposed to have sprung from the line of Aeneas, and thus from that of Venus. Among the Romans, traditions of descent remained vague even when written. Caesar's murderer, Brutus, was popularly supposed to be of the same family as an ancient Brutus, who had expelled the Tarquins, but no pedigree appears to have existed to substantiate the belief.

Among the northern nations that overwhelmed the western Roman Empire, belief in divine sonship was general. For Saxon rulers of the English kingdoms it was necessary to be descended from the god Woden.

Early written records. With the invention of writing, the oral became the written tradition. This occurred in Greece and Rome, where genealogies were recorded in poems and in histories. But genealogy did not at this stage become a science, because when writers dealt with it, they did so either incidentally in their narrative or because they were concerned with the family relationships of their gods.

Maximum length of genealogies

The historian Edward Gibbon's observation that "the proudest families are content to lose in the darkness of the middle ages, the tree of their pedigree" may be challenged in the light of recorded genealogies. The male line of Charlemagne has been traced to St. Arnulf, bishop of Metz, who died about 635. Several royal line descents are traceable to the 6th century, as, in England, is the tree of Louis Mountbatten, 1st Earl Mountbatten. The ancestry of Queen Elizabeth II goes to Egbert of Wessex (about 825), beyond him to Cerdic (*c.* 500), and, if another series of names is accepted, to Woden (an actual man later divinized by the Germans), in the 3rd century AD.

With the conversion of the barbarians to Christianity the recording of their regal traditions began. Examples occur in Ireland, Wales, and England. It was natural for the first chroniclers, who were mostly monks, to write down the oral pedigrees of the kings in whose realms they lived. Students of the Irish regal pedigrees are prepared to accept two or three generations before the time of St. Patrick (flourished 5th century AD) as genuine, and it is quite probable that name lists of the Irish kings are valid back to the 3rd century AD. Similarly, in Wales, the ancestry of the greatest Welsh families can be traced for a millennium. Among the Anglo-Saxons there were similar bardic pedigrees recorded by monastic scribes, and many of these might have survived but for the destruction of the Old English ruling class during the Norman Conquest. A regular feature of such old pedigrees recorded by monks was an attempt to link them with the genealogies of the Scriptures. In an Anglo-Saxon pedigree of great length—that of the kings of Wessex (the ancestors of Elizabeth II)—the line is thus traced to Sceaf, "a son of Noah born in the Ark." In the process of working out the connection between scriptural and regal genealogies, the monks adopted a reverse technique to that of the 4th-century-BC Greek mythographer Euhemerus; *i.e.*, they downgraded the old gods to human status.

From roughly 1100 to 1500, the emphasis of genealogists was on pedigrees of royal and noble lines. Claims to a throne, as with the dozen or so claimants to the Scottish crown after the death of Alexander III in 1286 and of his direct heir, Margaret the Maid of Norway in 1290, frequently involved genealogical trees. The truth was sometimes bent to suit some political end, but, on the whole, medieval European records are genealogically valid. This is because they were not primarily intended to supply genealogical information but to record land transactions, taxation, and lawsuits. The facts of family history are incidental and are therefore generally reliable. Exact dates of birth, marriage, and death are rarely given. A man is said to be of age "by Michaelmas 1330."

This period also saw the emergence of pedigrees of lesser folk. Land transactions involved claims in the local courts of the lords. Serfdom gave way to villenage; the latter involved so many days of labour on the lord's demesne and also the inability to move from the estate without the lord's consent. There was strong inducement for a man to prove that he was not a villein and for the bailiff to show that he was. In several parts of England, pedigrees of villeins or persons claimed as such have been worked out over periods of 100–150 years.

It was during the third period in European genealogical history that records that came to include everyone began. This period extends from 1500 to the present. As feudalism gradually gave way, new classes of citizens arose. In England the appearance of a powerful mercantile and business community was reflected in the growth of the middle classes, from which was continually recruited a new nobility and gentry. In turn, owing to the English rule of inheritance by primogeniture and the fact that unlike the

continental nobility English nobility has never extended beyond the reigning peer and his wife, the middle classes themselves continually received the younger children of peers and gentry. Two other factors leading to the proliferation of records were the enormous changes caused by the Reformation and the great reemphasis on individual religion and the desire of Renaissance monarchs to have more exact information about all their subjects.

Modern genealogy. Amateurs in the subject of genealogy are almost always actuated by the desire to trace their own family history. In the course of so doing they discover and work with general principles which apply to pedigrees other than their own, though records other than those applicable to their own case do not interest them. The professional genealogist is concerned not with one family but with many, and with the principles of genealogical research which arise from a wide study. As there are no university courses in the subject and therefore no degrees or other certificates of professional proficiency, the professional must be largely self-taught.

Qualifications of the professional genealogist

The disciplines required of a professional genealogist include a deep knowledge of the history of the country with which he is concerned and of its neighbours. National history determines the form of national genealogy, and genealogy can illuminate many aspects of national history that might otherwise remain obscure. The Wars of the Roses, for example, are hard to grasp unless genealogical trees showing relationships of the contestants are studied, and the course of the American Revolution is easier to understand when the links between George Washington and his compeers with the old English landed families who overthrew the Stuarts are comprehended. An understanding of the principles of law, especially of land law, the ability to decipher court hand or medieval script, an understanding of heraldry, and an intimate knowledge of the study of surnames and place names are also essential to the genealogist. Variations in surname spelling can be bewildering. The key is in the sound of the name, for a medieval scribe could not ask the illiterate person before him to spell his name.

The main task undertaken by professional genealogists is the tracing of pedigrees for clients, this being the staple of their work. Clients often consult genealogists when they wish to establish their family background, or, when having tried to trace it, they have come to a stop.

The writing of private family histories by professionals is very common. The material has usually been worked out by others who wish it to be checked and written by a professional.

Amateur genealogists, as already mentioned, are usually concerned only with their own families. The standard of amateur work varies with the individual, from the truly bad to the excellent.

Amateur genealogical work has increased greatly since 1945. In the United States there has been a long interest in the subject. The New England Historic Genealogical Society, the Augustan Society (based in California) and many state societies are of note. The Mormons (the Church of Jesus Christ of Latter-day Saints) have built up in Salt Lake City, Utah, a microfilm library of genealogical records from Britain and continental Europe, which is probably unequalled. In Canada, Australia, New Zealand, and South Africa the study of genealogy by private persons and by associations is growing rapidly. In England there is a Society of Genealogists, and there are corresponding bodies for Ireland and Scotland. In Denmark there is an International Confederation of Genealogy and Heraldry, which since 1954 has organized international congresses held in many European capitals at intervals of two years. In Czechoslovakia, by way of contrast, the national Genealogical Society has been dissolved, and in general it has not been feasible to obtain genealogical details from Communist countries, though it is probable that changes are now occurring in this respect. Jewish records are in a separate class. With the establishment of the state of Israel in 1948, a very great effort has been made to centralize information about the Jews of continental Europe under the care and direction of The Central Archives for the Study of the Jewish Peoples, in Jerusalem.

Main genealogical sources

In tracing family history, the worker follows certain rules. He works backward from the present. This is an elementary caution constantly put on one side by amateurs, who tend to trace forward from a person of the same name who may well be unrelated. As there cannot in the nature of things be a gap in a pedigree, no assumptions as to relationship can be allowed without very strong reason to accept them. Good and bad features in the ancestry have to be accepted: bastardy has to be allowed for, as well as regal ancestry by legitimate lines. An ancestor's wrongdoing must also be allowed for, though with passage of time this is usually taken in a romantic sense. Registration of birth, marriage, and death first became compulsory in England in 1837. Public records in most other Western countries began at varying dates in the 19th century. Census records are of great importance. They began in the United States as early as 1791; in Britain in 1801 (papers kept only from 1841); and even earlier in French Canada, in 1655–66. Parish registers began in England in 1538, though they are rarely preserved from that date. In most countries they begin later, but in Spain the oldest extant is dated 1394, and there are 1,636 parishes having records prior to 1570. In England, Nonconformist records have been kept by various bodies, and many are now held officially at Somerset House, London, or the Public Record Office. In America the settlers were generally trying to get away from established church and controlling state. They were vigorous individualists who kept careful records of their lives and of the organization of their new communities. Examples of detail in New England records can be seen in many of the 1,600 pedigrees contained in *American Families with British Ancestry* (500 page suppl. to Burke's *Landed Gentry,* 1939), in A.M. Burke's *Prominent Families of the United States of America,* and in the many volumes of family history produced in the United States.

Wills are of the utmost importance as a source of genealogical information. Ships' lists of passengers are useful in supplying dates for immigrant ancestors' departure for the New World, but since they do not indicate place of origin, but only the port of departure, the original habitat must be sought elsewhere. Without knowledge of the ancestor's place of origin in the homeland it is useless in almost all cases to attempt a search.

With the aid of the type of record mentioned above, and with help from family Bibles, tombstones, and plaques and brasses in churches, it is as a rule possible for a person of English antecedents to trace some 250–300 years of ancestry. Before the 17th century, everything depends on the social position of the ancestors. Tax records, lawsuits, and purchases and sales of land are the chief sources for tracing a family before 1600. The Pipe Rolls extend from the reign of Henry II (1154–89) to that of Queen Victoria (1837–1901), with an interrupted beginning also in the time of Henry I (1100–35). Monastic records are of great importance as showing grants or ownership of land. The pleas of the crown deal with suits at law and contain much detail about families. There are many Rolls besides those of the Pipe that give a great deal of incidental genealogical information. *Inquisitiones post mortem* show the position of an heir; *i.e.,* his age and other details. As the centuries are passed, the numbers of those who can prove a descent by the male line dwindle, until by the time of the Norman Conquest scarcely half a dozen pedigrees can be traced in the male line for either Saxon or Norman.

Regarding deposition of public records, two principles have been followed by archivists: that of centralization and that of diffused local holdings. The former has many obvious advantages and was adopted in Scotland and in Ireland. It has one disadvantage—destruction of the records at one stroke. This happened in Dublin on April 13, 1922, when Irish factions fighting with each other burned most of the Irish records. The second system, by which records are stored in a number of depositories, prevails to a considerable extent in England. Although the Public Record Office, Somerset House, and the British Museum library are places of centralized record, the parish registers remain outside them, scattered in numerous parishes or county offices. County records contain masses of material not to be found in London.

From the 16th century there has been an increasing accumulation of written material, which deals either exclusively or incidentally with genealogy. William Camden (1551–1623), a learned English antiquary and historian, did much to raise the standards of genealogical research. He was the first English writer on surnames, and his work was not resumed for nearly 200 years. Sir William Dugdale, a younger contemporary of Camden, made a beginning with his *Antiquities of Warwickshire* to the great output of county histories written between the 17th and 19th centuries. The revolution in modern genealogy was the application to its study of canons of historical and literary criticism formulated in Europe from 1800. Their application to genealogy was fairly late, as is illustrated by the fact that the 19th-century English historian Thomas Macaulay, critically perceptive in most other spheres, accepted what amounted to family myths as true genealogy. Later writers, including J.H. Round, W. Farrer, C. Lewis Loyd, C.T. Clay, and the editors of the *Complete Peerage*, are of the greatest importance. (L.G.P.)

PALEOGRAPHY

Paleography (from Greek *palaios*, "old," and *graphein*, "to write") is the study of ancient and medieval handwriting. Precise boundaries for the study are hard to define. For example, epigraphy, the study of inscriptions cut on immovable objects for permanent public inspection, is related to paleography. Casual graffiti, sale or election notices as found on the walls of Pompeii, and Christian inscriptions in the Roman catacombs are likewise part of paleographical knowledge. In general, however, paleography embraces writing found principally on papyrus, parchment (vellum), and paper. Today, paleography is regarded as relating to Greek and Latin scripts with their derivatives, thus, as a rule, excluding Egyptian, Hebrew, and Middle and Far Eastern scripts. It is closely linked with diplomatic (see above), the study of forms in which official and private documents are drawn up.

The scientific study of Latin paleography (and diplomatic) dates from 1681, when the French monk Jean Mabillon published *De Re Diplomatica*, the first textbook on the subject, while his compatriot Bernard de Montfaucon performed a parallel service for Greek paleography in his *Palaeographia Graeca* in 1708.

The primary task of the paleographer is to read the writings of the past correctly and to assign a date and place of origin. Close acquaintance with the language of the text is a prerequisite. Help in dating is offered by changes in styles of handwriting and variations from area to area. Abbreviations in texts likewise help in dating and localization.

Types of writing materials. A paleographer must be familiar with writing materials. Any smooth surface able to accept writing has served in the past, notably, pottery fragments, animals' shoulder blades, slabs of wood, bark, cloth, and metal.

Papyrus

The great writing material of the ancient world was papyrus, in use by 3500 BC. In preparing the surface, strips taken from the papyrus reed (*byblos*) growing in the Nile Delta were laid side by side, while other strips were laid across at right angles, and the whole impregnated with paste. After treatment a fine smooth surface was obtained. Much of the administration of the Roman Empire depended upon papyrus, in the same way that modern bureaucracies depend upon paper. Warfare and a damp climate resulted in an almost total disappearance of papyrus from Europe, though the dry (if war-scarred) sands of Egypt have preserved vast numbers of documents. Papyrus was imported into Europe from Egypt even after the fall of Rome. Chance survivals include charters of Merovingian kings in France (7th century) and business documents (5th–10th centuries) at Ravenna, the old administrative capital of the late Roman Empire.

The other great ancient writing material, still in occasional use today, is parchment, or vellum, the terms being often used interchangeably. Vellum is a term usually applied to skin from a calf (*cf.* veal, *veau*), while parchment is an expression often applied to sheepskin or goatskin. The word parchment is derived from Pergamum in Asia Minor, the ancient centre of its manufacture.

Both papyrus and parchment were expensive and were replaced for everyday use by wax tablets corresponding to today's notebook. Tablets made of wooden blocks were hollowed out and filled with melted, often black wax. Notes were made in the hardened surface. Even documents of permanent significance, such as property conveyances, were made on wax tablets.

Because ancient writing materials were expensive, they were often reused. Papyrus presented difficulties, for ink soon bonded itself firmly into the surface. Parchment could be more readily reused, because it is tougher and can be washed or scraped clean. Many medieval monks, when short of writing materials, took ancient books to pieces, cleaned off the leaves and used them again. The original script can often be brought out under ultraviolet light. Parchments thus cleaned and freshly inscribed are called palimpsests (Greek *palin*, "again"; *psēstos*, "scraped").

Paper is the third great writing material. In use in China at a remote period, it was employed extensively in the Arab world by the 9th century. Not in common use in Europe until the 14th century, it took over the name of the half-forgotten papyrus.

In the early classical world the standard form of book was the papyrus roll, commonly called *biblion*, taking its name from the material of which it was made. It consisted of papyrus sheets pasted edge to edge with a slight overlap. The text was set out in columns, drawn up at right angles to the edge of the rolls, and started at the left. The reader unrolled as he went along and at the conclusion was obliged to reroll the book. The roll was an inconvenient form of book, difficult to consult, which probably accounts for the inaccurate quotations found in early literature, caused by an author relying on his memory rather than troubling to unwind a long roll. By the time of Christ, a new form of book was coming into fashion, the codex, or book in the shape in which it is known today. The codex is almost always of parchment, since papyrus cracks when folded. The codex seems first to have been used for notebooks or account books, the conservatism of booksellers and readers ensuring the survival of the roll for centuries. The Christians popularized the codex, using it for the Gospels.

Form and binding of texts

Various instruments have been used for writing. The early Egyptians used a slender rush. From about 300 BC the thicker reed pen was used. The reed was in general use in the Greco-Roman world. Metal pens, copied from the reed, were also employed. For wax tablets a stylus was used, made of wood, bone, ivory, iron, or bronze. In many northern European areas, where reeds suitable for writing purposes are not indigenous, the feather (*penna*) became the main writing instrument. It was usually stripped of its vanes and the quill alone used.

Writing implements

Lead, used in classical times for ruling guidelines in manuscripts, was used extensively in the Middle Ages for rough notes and annotations in the margins of books.

Ink has been prepared in a variety of ways. In classical times the black discharge of cuttlefish was used, as well as concoctions of soot and gum. In the Middle Ages oak apples were steeped in water with "vitriol" (ferrous sulphate) to produce ink.

Analysis of texts. The essential skill of a paleographer is the ability to recognize the numerous styles of handwriting prevalent in different ages and places. Most European scripts descend from Greek and Roman capital letters, but variations are enormous. It is a European convention that writing starts on the left at the top and works line by line down the page. An eccentricity known as boustrophedon (from Greek *boustrophēdon*, "following the ox furrow"), whereby alternate lines are written backward in mirror writing, occurs chiefly in very ancient inscriptions.

The Greek and Latin alphabets existed originally as capital, or majuscule, letters. The ancient Greek alphabet, as developed in chiselled inscriptions on stone or marble, served without much modification as the alphabet used in literary works written on papyrus rolls. This script, found in the oldest surviving Greek literary papyri of *c.* 300 BC or earlier, gave way to more rounded and elegant

forms, probably developed in the Greek literary circles of Alexandria. Cursive scripts that were easier to write were developed for everyday use, for business, and to record the acts of the great bureaucracy of Egypt, where the Greeks settled in large numbers. The Greek cursive script and the formal book script greatly influenced each other, as can be seen from a vast series of cursive documents dating from the 4th century BC for about 1,000 years. Because so much material survived, early Greek cursive can be better studied than its Latin counterpart. In Greek cursive manuscripts the everyday life of ordinary people becomes a reality: they pay or fail to pay taxes, buy or sell houses, and harass civil servants with awkward demands.

A very rough division in Greek paleography may be made at around AD 300. The earlier age is called the papyrus period; and the later, the parchment or Byzantine (or Christian) period. The division, however, is imprecise, for parchment was used well before and papyrus long after this date. The change from papyrus to parchment is signalized by three great monuments of paleographical studies, the Vatican, Alexandrine, and Sinai Bibles, all on parchment and in codex form.

An alphabet of small, or minuscule, letters developed gradually and was in use by the 8th century. Numerous abbreviations exist in Greek manuscripts, though never so many as in Latin. Accents, an additional complexity, were not systematically applied before the 7th century AD.

The ancient Latin alphabet of capitals (*quadrata*) is found in numberless inscriptions in stone and marble all over the Roman world. How far this alphabet was used for writing books is uncertain, because, though excellently adapted for incision, it is difficult to write. Some specimens of handwriting in *quadrata* do exist, such as 4th- or 5th-century copies of Virgil, but scholarly opinion largely regards these as abnormal productions. By the 1st century a handsome Latin alphabet existed, called rustic, based on the use of a broad pen or brush. Rustic was used for public inscriptions on walls, as in the sale and election notices found at Pompeii. Although specimens are scarce, it is likely that books were extensively written in this hand in classical times. By the 4th century another Latin alphabet existed, the script known as uncial, in the nature of a rounded form of *quadrata*. Uncial survived the fall of Rome and from it developed half-uncial, the ancestor of the small letters in use today.

Styles of writing

The stately Roman scripts, *quadrata*, rustic, or uncial, were not used for everyday purposes, and, as in the case of Greek, a cursive, rapidly written hand arose in which letters and business documents were inscribed. This hand is found in graffiti on Pompeian walls and in wax tablets. After the disintegration of the empire, Roman cursive became the ancestor of regional hands in what are now Spain, France, and Italy.

During the flowering of Christianity and art in Ireland (*c.* 500–*c.* 1000) a beautiful "insular" script developed, which found its way into England. There, two streams of influence commingled, for from 597 Christian missionaries arrived from Rome and brought in books in uncial script. Both scripts prospered in England, though insular gradually superseded uncial.

The most successful of all scripts proved to be Caroline minuscule, which takes its name from the emperor Charlemagne (died 814), patron of scholars and scribes, under whom the script was developed. Despite its inherent superiority and clarity, it did not predominate over regional scripts until the mid-12th century, and the local hand of southern Italy (Beneventan) maintained itself for much longer.

In the 12th century, Caroline minuscule, which had undergone moderate developments, started to display more obvious changes. It compressed laterally, while its rounded strokes became stiffer and straighter as it was converted into the so-called Gothic hands—very angular in northern Europe and more rounded in Italy. A revulsion against Gothic took place in scholarly circles in Italy in the 14th and 15th centuries, and a return to models based on Caroline minuscule took place. This revived hand, called Humanistic because humanist scholars used it, was adopted by 15th-century Italian printers, whose type faces ultimately triumphed over the Gothic. (This encyclopaedia is printed in a type scarcely modified from Caroline minuscule.) Meanwhile, Caroline and Gothic scripts had produced cursive hands for quick everyday use, as in the case of the ancient Greek and Latin alphabets. These cursive scripts were used for the vast mass of business documents written in the Middle Ages.

Cursive scripts

Abbreviations are the principal problem confronting paleographers. They were extensively used in Roman times by lawyers to avoid repetition of technical terms and formulas. Abbreviations fall into two classes, suspension and contraction. Suspension, omission of the end of a word and indication by a point or sign, was used in Roman public inscriptions—*e.g.*, IMP.(ERATOR), CAES.(AR). Contraction, the omission of letters from the middle of a word and replacement by a sign or some other device, was common among Greek-speaking Jews, who contracted certain sacred or revered names, such as God, Lord, Israel, or David, as a mark of veneration. The Christians followed the practice by contracting their sacred names, such as Jesus and Christos. The great increase in use of abbreviation as a means of saving time and material dates from the 12th century, but some contraction signs are of high antiquity, such as the sign 7 for *et* (Latin, "and"). Some are quickly written versions of letter groups, such as "÷" for *est* (Latin, "is"), the top dot standing for e, the bottom dot for t, and the stroke being a long or f-shaped s, fallen on its side. The letter r is often omitted, the adjacent vowel being written above the line, as c°a for *carta* ("charter"). In works for semilearned readers, such as romances, abbreviations are often few, but books produced for the learned, such as university textbooks, are heavily loaded with abbreviations. The number of signs and devices in use by the end of the Middle Ages was enormous. More than 13,000 are listed in the standard work, Adriano Cappelli's *Lexicon Abbreviaturarum* (1912).

Abbreviations

Dating of books and documents also offers problems. Even when a precise date is given, the dating system of a given time and a given area must be checked because the year began at different times in different territories, and there were even variations in the same country. Calendar reforms initiated in 1582 by Pope Gregory XIII, for example, were not adopted in Protestant England until 1752. If a year of a monarch's reign is given as a date, it is necessary to determine whether the reign is counted from his accession or coronation. Moreover, few books were dated, the dated title page being nonexistent in medieval works, though sometimes a final paragraph, the colophon, supplies a date with the scribe's name and place of work. Important documents, such as English 12th-century royal charters, are undated.

In the absence of dates, inferences are drawn from handwriting, use of abbreviations, and internal evidence. Caution must be used, however; for an elderly scribe may be using a hand learned over half a century before, and work in the same scriptorium or office as a young clerk anxious to show off all the latest tricks and flourishes. Some styles lasted a long time: Caroline minuscule lasted for more than three centuries. Certain kinds of books, such as liturgical volumes, were produced in a highly stylized form for generations, and thus it is often difficult to provide a close date for a late medieval missal (with standard illustrations and marginal decorations) in a mechanical Gothic hand.

Internal evidence must be weighed carefully. A given historical event noted in a chronicle will provide an earliest possible date, unless the entry is an interpolation. Evidence for some legal practice or liturgical usage is no safe guide, for legislation on the subject by a king or a pope may merely be ratification of a long-standing practice.

A paleographer must get to know his scribes, for their mannerisms can be highly informative. Nearly 50 different scribes have been distinguished in the English royal chancery in the period 1100–89. First-rate scribes, such as notaries public, provide much information about themselves, giving their names, notarial signs, and information on their authority to act. Even anonymous clerks, who drew up innumerable property conveyances, can be identified by their script over a period of years, and their career

Internal evidence

can be traced through developing, mature, and deteriorating handwriting, thereby offering dating evidence.

The provenance (origin) of many manuscripts can be immediately recognized because certain centres developed individual styles. Papal and royal chanceries issued documents of easily identifiable origin, while many monastic scriptoria—for example, that of Canterbury Cathedral in the earlier 12th century—had a virtually private handwriting.

Textual corruptions are another obstacle to correct elucidation. A legal document is certain to have been checked at the time of writing, but one cannot be sure in the case of a literary, philosophical, or theological text. Scribes were fallible, and, if there are no signs of any corrections in a text, then it probably embodies inaccuracies. A popular book, such as Chaucer's works, exists in large numbers of manuscripts, and many manuscripts produce variant readings. If a scribe made a mistake in copying, future scribes using his version are likely to reproduce the error and add others. Sometimes the same muddled passage in a group of manuscripts of a given author can be traced back to damage in an earlier copy, say a section eaten by rodents or impenetrably stained. Whenever copyists worked from different and faulty originals, various copies tend to fall into families. A paleographer must bring together various readings in families and decide which is the best reading.

Sometimes a scribe, set to work because he could write a fine hand, did not necessarily possess much knowledge of the language. Such a scribe faced with a text heavily loaded with abbreviations would usually make nonsense of it. Occasionally, a particularly stupid copyist, faced with a master copy in two columns of writing, would copy straight across the top line, then across the second. When he used a different number of words per line, the text was reduced to unintelligibility. In Greek and Roman times there was the difficulty that texts were written continuously, without space between the words. Copyists misread passages. For instance the historian Tacitus reported that some tribesmen went off to guard their own property: ADSVATVTANDA (*ad sua tutanda*). Some copyist thought "Suatutanda" was a place and this ghost name was perpetuated in geographical works. Later medieval Gothic hands presented a forest of vertical strokes called minims. The letter *v* rendered as *u* made two strokes, while *i* was often left without a dot or at best with a faint hairline, often misplaced. The group of letters *ium* could be read, as *uim, uiui, niui, mui, miu,* with many other variations. Accordingly, minim corruption, confusion of vertical strokes, is a term constantly heard in paleographical circles.

Latin and Greek are inflected languages in which the same case and tense endings constantly occur, offering scope for error. Moreover, in biblical, theological, or philosophical texts, the same words abound. For example, in one place in the Gospel According to John there occurs the passage:

> Verba quae ego loquor
> vobis a me ipso non loquor
> pater autem in me manens . . .

("The words that I speak . . ."). The eye of a sleepy scribe might slip from the first *loquor* to the second, whereupon he would go on copying at *pater autem,* leaving out the second line altogether, a common type of error known as *homoioteleuton* ("like ending").

Because of the lack of surviving specimens, it is difficult to assess book decoration in classical times, but apparently it was very limited. In the later centuries of the Roman Empire, however, book illustrations were not infrequent.

Decorations

The narrative material in the Bible encouraged illustration. The Irish were foremost in applying decoration to the text in the form of elaboration of capital letters, producing such masterpieces as the Book of Kells (late 7th century), in which Celtic imagination and artistic sense ran riot in elevating the book to an object of outstanding beauty. Some of the greatest creative talent of the Middle Ages was lavished upon books, especially upon those used in worship, such as Bibles, psalters, and missals. When a book cannot be assigned either a date or provenance upon the appearance of the text alone, its style of illumination will often direct the paleographer to a certain monastery in which the carving on capitals or wall paintings may contain the same motifs.

Because of the immensely high prices of manuscripts, the question of forgery naturally arises, but it is safe to say that no modern forgery could survive for a moment. A convincing imitation of ancient script is virtually impossible, while the papyrus, parchment, or paper on which it would be written could not stand up to modern scientific inspection. Anything of recent vegetable or animal origin fluoresces brightly under ultraviolet light, to name but one test. William Henry Ireland (died 1835), the Shakespeare forger, used flyleaves from 16th-century books, but his handwriting and non-Shakespearean language gave him away. A modern would-be forger must either copy an existing work, which, in the present state of art history and paleographical study, would be immediately recognized, or be prepared to invent medieval subject matter.

Forgeries

There was fabrication of documents in medieval times on a considerable scale. A monastery might find itself in possession of estates held since remote antiquity but without any title deeds. When some powerful monarch made difficulties, there was a strong inducement to produce the required ancient-looking documents. The borderline between justifying legitimate possession and culpable attempts to gain extra territory or privileges, however, is ill-defined. Monks occasionally descended to falsifications of title deeds and charters of exemption. About 1125 a monk of Soissons on his deathbed confessed to a career of professional forgery for gain and admitted fabricating charters for various monasteries, including Westminster Abbey. Early forgeries, however, give themselves away through such inconsistencies as mentioning bishops of nonexistent sees or embodying legal phrases that came into use generations later or bearing seals when seals were not yet appended to documents.

The modern paleographer has great technical aids: photography since the 19th century and colour photography in the 20th. Ultraviolet light brings out faded handwriting. Microfilm makes the contents of a volume in a far-distant repository available quickly and cheaply. (W.G.U.)

SIGILLOGRAPHY

Sigillography is the study of seals. A sealing is the impression made by the impact of a hard engraved surface on a softer material, such as clay or wax, once used to authenticate documents in the manner of a signature today; the word seal (Latin *sigillum;* old French *scel*) refers either to the matrix (or die) or to the impression. Seals are usually round or a pointed oval in shape or occasionally triangular, square, diamond, or shield-shaped.

Medieval matrices were usually made of latten—a kind of bronze—or of silver. Ivory and lead were occasionally used, gold very rarely. Steel was used from the 17th century. Matrices could include intaglio gems. The usual material for the impression was sealing wax, made of beeswax and resin, often coloured red or green. In southern Europe, notably in the papal Curia, lead and occasionally gold were used. Shellac, the wax used today, was introduced in the 16th century.

Purposes of seals

Seals were used to establish the authenticity of such documents as charters and legal agreements and for the verification of administrative warrants. In southern Europe, early medieval documents were drawn up by notaries and authenticated with their written *signa,* but this never replaced seals in northern Europe. Forgeries were manufactured as early as the 12th century, indicating how important seals had become. From that time, also, seals were used to close folded documents and thus to guarantee their secrecy. Seals were also used to affirm assent; for example, by a jury. Under the Statute of Cambridge (1388), sealed letters were used in England for the identification of people and their places of origin.

Sigillography is used to assist other historical studies. Many impressions have survived from the medieval period. Those attached to documents are most valuable, because the documents may date their use precisely and the seal may confirm the documents' authenticity. Unattached seals may still provide useful evidence from their inscription or design. Fragmentary seal impressions are often

difficult to interpret. Fewer matrices have survived and often without related impressions. Seals often reflect the taste of the owner. They provide evidence for changes in fashion in both secular and ecclesiastical costume and for the development of armour. Seals indicate heraldry before the earliest rolls of arms and are an original source for armorial bearings, which thereby enables the historian to trace the distinctions or alliances between various families and so contributes to genealogy. Craftsmen's seals often display tools connected with trade. Depictions of towns, churches, castles, and monasteries, although conventional, can often aid the architectural historian. Seals can also be used profitably for studying ancient ships, particularly their shape and details of masts and rigging. The main difficulty in studying seal designs is that they were often conservative, especially since seals were often replaced with an exactly similar design.

(J.Ch.)

Seals in antiquity. Seals with designs carved in intaglio were used throughout antiquity. They were of two main types—the cylinder and the stamp. The cylinder first appeared in Mesopotamia in the late 4th millennium BC and continued to be used there until the 4th century BC. It was also widespread in Elam, Syria, and Egypt (3rd millennium BC) and in Cyprus and the Aegean (2nd millennium BC). Stamp seals preceded cylinders, first appearing in Mesopotamia in the 5th millennium BC and developing over a period of about 1,500 years until largely replaced by the cylinder in the 3rd millennium. Early stamp seals were also used in Iran, northern Syria, and southeastern Anatolia during the 4th and 3rd millennia; in Anatolia their use was continued in the 2nd millennium by the Hittites. In Mesopotamia the stamp seal gradually came into use again in the 8th–6th centuries, effectively replacing the cylinder by the 3rd century BC. In Egypt the scarab largely replaced the cylinder seal early in the 2nd millennium BC and continued as the main type until replaced by the signet ring in Roman times. In the Aegean, various types of stamp seals were used throughout the 2nd and much of the 1st millennium BC, until in Hellenistic and Roman times the signet ring became dominant.

The uses of ancient seals are known from textual references and ancient sealings, both on lumps of clay and on documents found in excavations. In historical times most prominent citizens, including women, carried their own seals. That the rank or office of the owner was often included in the inscription indicates that many of these may have been official seals. Kings had their own seals, and high officials could hold the king's seal as a mark of delegated authority.

Uses of ancient seals

Seals came into use before the invention of writing for the securing of jars, bales, bags, baskets, boxes, doors, etc., and this use continued throughout ancient times. The method was either to shape clay over the stopper or lid or to make a fastening with cord and place clay around the knot and then impress it with the seal.

The sealing of written documents, of which the two major ancient classes were clay tablets and papyrus scrolls, became regularly established in the latter part of the 3rd millennium BC. The clay tablet was the main vehicle of writing in Mesopotamia, where cuneiform was used into the Christian Era, and this method spread to Elam, Anatolia, Syria, and Egypt; clay tablets were also used in the Aegean Bronze Age. The tablet was inscribed while the clay was soft, and seal impressions were applied at the same time.

The main kinds of sealed cuneiform documents were contracts, accounts, and letters. Contracts were sealed by the contracting parties and by witnesses and commonly encased in clay "envelopes," on which the text was either repeated or summarized and the seals again impressed. Two special kinds of contract were the royal grant to a subject, impressed with the royal seal of the grantor, and the treaty between nations, a number of examples of which have been recovered, some of them bearing impressions of the seals of the royal contracting parties. Account tablets were sealed to authenticate the transfer of goods. A letter was often encased in an "envelope" and the sender's seal impressed on the outside to identify him to the recipient and, in the case of official letters, to authenticate any commands contained in the letter.

In Egypt, papyrus documents may be assumed from soon after 3000 BC, but surviving evidence dates mostly from the latter part of the 3rd millennium onward. The method of sealing a papyrus document was to roll it into a tube, tie a strand or cord around the centre, and seal a clay lump over the knot. This method continued into the Christian Era, from which time a great number of Greek papyri have survived in Egypt. It was not until the 1st millennium BC that this kind of document, including by then leather and parchment, came into wide use outside Egypt. The spread of this kind of document, on which the space for a seal was small, probably played some part in the gradual replacement of the cylinder by the stamp seal.

No documents or sealings have been discovered from ancient India, but the still undeciphered inscriptions on the seals may include personal names, perhaps of merchants, who could have used the seals in much the same ways as their Near Eastern contemporaries, with whom they are known to have had commercial contacts.

Since seals were used throughout ancient times and are sufficiently durable to have survived in very large numbers, they form one of the few classes of ancient objects in which a continuous development can be traced. The great majority bear artistic representations, so their chief value is for art history, but, since these include details of environment (plants, animals), equipment (plows, chariots, musical instruments), or dress, they also contribute to cultural history.

Further information is provided by the inscriptions on seals. The existence of rulers known only from king lists may sometimes be confirmed by the discovery of their seals, and in some cases rulers are known only from their seals, which, because they often mention the names of their fathers, the cities that they ruled, and the chief gods that they served, form a valuable historical source. The assembling of tablets and sealings bearing the impressions of private seals can contribute to the reconstruction of business archives and the destinations of traded goods, thus providing valuable material for economic analysis, and far-flung trade contacts can be deduced from foreign seals in excavations (*e.g.*, Indus Valley seals in Babylonia). Personal names are an important source for ethnic analysis, and inscribed seals, because they often name the owner's father and even grandfather, provide material for this as well as for genealogical reconstruction. (T.C.M.)

Medieval European seals. The connection between Roman and medieval seals lies in the use of seals in the chanceries of the Merovingian and Carolingian kings. Many Ottonian seals had busts of the emperors. Royal seals of medieval type, with the ruler enthroned and bearing his insignia, appear from the 11th century. The use of seals by bishops and nobles became usual at this time and was widespread by the 12th century. By the 13th century, seals were used by all classes, including small landowners; and, by the 14th century, simple seal matrices could be bought ready-made.

Spread of sealing in medieval times

The quality of engraving varied greatly. Some delicately designed seals date from the 12th century, such as the silver seal of Isabella of Hainaut, queen of France in 1180–90. The silver equestrian seal of Robert Fitzwalter is a notable example of the 13th century, the period of the finest seal engraving.

The names of several engravers of medieval seals are known: for example, Luke, who engraved the seal of Exeter, and Walter de Ripa, who engraved the first great seal of Henry III of England.

Forms of medieval seals. Seal matrices may be single or double, thus producing an impression on either one or both sides of the wax. Single matrices, the older type, often have a ridge along the back and end in a loop. Double matrices, known from the 11th century onward, are flat, with two to four projecting lugs pierced with holes in which vertical pins keep the halves aligned.

Sealing both sides of the wax makes detaching the seal more difficult, and so in medieval times the reverse was often sealed by a counterseal for greater security. The official seal of an institution was often countersealed by

the seal of an official, such as a town by its mayor. Single seals were often fitted with a handle; the most common type was a six-sided cone ending in a trefoil. In some matrices the centre screwed outward, enabling the device to be used without a legend. On many seals the back was marked with a cross to indicate the top.

Seals could be either applied to the surface of a document or appended from it by a strip of material. Application was the earlier system, although papal bulls were always appended. Appended seals appeared in England in the 11th century and in France in the 12th; seals were appended either on a tongue of parchment cut across from the bottom of the document or on a tag of parchment, leather, or silk inserted through a cut in the document. Some documents had many seals. Seals were often protected by woven bags or by boxes of wood, metal, or ivory known as skippets.

The legend

The legend, often abbreviated, usually declared the name of the owner or institution; it often began with a cross and the word *sigillum*, followed by the name in the genitive case. Latin remained in fashion for inscriptions, though English and French are occasionally found from the 13th century, more frequently on personal seals. On English seals roman capitals were used in the 11th and 12th centuries and Lombardic ones in the 13th and early 14th centuries. Black letter (Gothic script) was first used in England in the 14th century and was quite popular in the 15th century, although Lombardic often continued for capitals. Roman capitals reappeared in the 16th century.

Royal and official seals. The great seal, or seal of majesty (a round seal showing the seated ruler with the royal insignia), first appeared in Europe on the seal of the emperor Henry II of Germany (ruled 1002–24), in France on the seal of Henry I (ruled 1031–60), and in England on the double seal of Edward the Confessor (ruled 1042–66). The seal of William I of England (ruled 1066–87) had the King on one side and an equestrian figure on the other. The kings of France adopted double seals under Louis VII (ruled 1137–80).

The development of lesser royal seals can be illustrated by the growth of English government. Deputed great seals were used for the major legal courts and for France, Ireland, and Wales. The expansion of the kings' affairs caused the addition of smaller, more personal seals, such as the signet. The Chancery did not control these seals, and this freedom led to the evolution of autonomous offices. The privy seal appeared early in the 13th century in the custody of the clerks of the king's chamber. It was soon transferred to the wardrobe clerks, and gradually its importance increased until by the early 14th century the keeper of the privy seal was the third minister of state. The keepership gained further prestige in midcentury, when the great seal was entrusted to the keepers who went abroad with Edward III. As the privy seal grew in importance, the king preferred another small seal for authenticating correspondence and warrants. Under Edward II (ruled 1307–27) there was a secret seal distinct from the privy seal. By 1400 the signet, as the secret seal was then called, was in the charge of the king's secretary. The signet rather than the privy seal became the originating force in administration, and from 1540 there were two secretaries, each with two signets. The privy seal and signet seal were both single armorial seals.

Royal officials had their own seals. Circular admirals' seals, dating from the late 14th century to the 17th century, include a fine group of 15th-century bronze matrices. The seals show ships in great detail, with the sails displaying the arms of the admiral.

Bishops' and monastic seals

Religious seals. The principal episcopal seal was the seal of dignity, always a pointed oval. From the 11th to the 14th century it usually depicted the standing figure of the bishop, from the 13th century with a canopy above him. In the mid-14th century the standing figure was often replaced by a saint or a religious scene, with the bishop praying beneath—a form that had been used earlier on episcopal counterseals. The seal of Thomas Arundel, archbishop of Canterbury (1396), depicts the martyrdom of Becket in the centre of an elaborate series of niches, with the archbishop below.

Monastic seals, usually double-sided and of high quality, normally show the buildings of the monastery, religious scenes, or the patron saint. They were distinct from abbots' and priors' seals, which were similar to those of bishops. Notable was the elaborate four-part matrix of Boxgrove Priory (mid-13th century). The seal of Merton Priory (1241), considered the finest English medieval seal, had the Virgin and child on one side with St. Augustine of Hippo on the other.

Papal bulls were doubled-sided lead seals appended to the document on strings. The earliest known is that of Deusdedit (reigned 615–618). The usual design, with the head of the Apostles Peter and Paul on one side and the pope's name on the other, first appeared under Paschal II (reigned 1099–1118). Although this style of portrayal of the heads was changed in the Renaissance, the design has not been altered.

Town seals. The possession of a common seal was an important part of a town's independence. Town seals were almost always round and often double. Many towns still possess their original matrices. The earliest in England date from around 1200, when many towns received their charters. The seal of Exeter has been dated to *c.* 1180. Maritime towns often depicted a ship with a furled sail; inland towns often showed the guildhall or the town itself. The seal of Rochester depicted the Norman castle within a wall. Counterseals often bore the figures of saints. Later medieval town seals were less common and beginning with the 14th century were often single seals.

Commercial seals. Commercial seals in England were considerably increased during the reign of Edward I, when double-sided bronze dies occurred for various customs, subsidies, and the delivery of wool and hides. Their obverse displayed the arms of England, and the reverse had the same device without the shield. The seals of merchants and craftsmen often displayed either merchants' marks or tools connected with their trade.

Personal seals. The earliest class of personal seals was that of greater barons of the 12th century, who used the device of a fully armed equstrian knight. Their shields often provide the earliest evidence for the use of heraldic charges. Some greater barons used a double seal with an equestrian obverse and an armorial reverse. The most usual type of personal seal was a single seal with the arms of the owner. Women's seals were usually pointed ovals and showed the lady standing, sometimes between shields. Nonheraldic personal seals displayed a variety of devices, such as stars, fleurs-de-lis, armorials, and religious subjects. The inscription sometimes indicated the owner, although it may simply have related to the device, as on those that bore the device of a squirrel and the inscription "I krack nuts."

Modern use of seals. The use of seals declined as the use of signatures grew. Personal fob seals were in fashion from the 17th to the early 19th century; often they were gems in gold settings that were carried in the fob or breeches pocket and were used to seal folded private correspondence before the envelope was introduced. States and institutions continued to use seals for the formal ratification of their acts, but few of these seals maintained the medieval vigour of design. (J.Ch.)

Chinese and Japanese seals. The private seals used in China (*t'u-chang*) and Japan (*ingyō*), commonly square and reading merely "seal of so and so" (XX *chih yin*), served as a confirmation of signature or a sign to be verified but have not the legal status of a signature. They are made of ivory, wood, or jade. Used by artists and collectors to mark their paintings and books, there is hardly a limit to their fanciful designs and phrasеology. A man might own scores of seals, using his many sobriquets, especially those suggesting unworldly and rustic tastes. A seal is impressed in red ink—made of cinnabar in water and honey or suspended in sesame oil, hempseed oil, etc.—held ready on a pad of cotton or moss. The characters most often appear in line, but they are sometimes reserved against the inked ground.

The first record of a seal in China is from 544 BC. Actual bronze seals survive from the 5th century BC, and the practice of sealing must be some centuries older. The emblematic characters cast on Shang dynasty bronze vessels

(13th–11th century BC) imply the use of something like a seal for impressing on the mold. The royal seal and other seals of high office were termed *hsi*; other seals of rank and appointment were *chang*. The imperial *hsi* (called *pao* beginning in the T'ang period, AD 618–907) was traditionally large and square, made of jade or ivory. The most famous one belonged to Shih Huang-ti (ruled 221–209/210 BC); it had as its knob a one-horned dragon and is fabled to have been handed down to the present day.

"Small" and "great" seals

The official and, no doubt, the personal use of seals began in Japan with the copying of Chinese institutions in the 7th century AD. Both in China and in Japan modern seals generally employ the "small seal" character (*chuan shu*), the "great seal" character being reserved in the past for the ruler and high officers. To the historian the importance of the Far Eastern seals is greater in the earlier periods, and in China they yield more information than in Japan. Thus, seals recovered archaeologically throw light on government appointments made in the Han period, particularly in the reign of Han Wu Ti (140–87 BC), when they were tokens of rank given to internal officials and some external client rulers. Gold seals of the "King of the Han Wei-nu county," found near Fukuoka in 1784, and that of the "King of Tien" excavated near K'unming in 1956 have implications of this kind. But in post-Han times the seals have served little if at all as primary historical documents, and in writings on East Asia it is chiefly the art historian who appeals to their testimony in authenticating paintings and calligraphies.

(W.W.)

TEXTUAL CRITICISM

Textual criticism defined

The technique of restoring texts as nearly as possible to their original form is called textual criticism. Texts in this connection are defined as writings other than formal documents, inscribed or printed on paper, parchment, papyrus, or similar materials. The study of formal documents such as deeds and charters belongs to the science known as "diplomatic"; the study of writings on stone is part of epigraphy; while inscriptions on coins and seals are the province of numismatics and sigillography. Textual criticism, properly speaking, is an ancillary academic discipline designed to lay the foundations for the so-called higher criticism, which deals with questions of authenticity and attribution, of interpretation, and of literary and historical evaluation. This distinction between the lower and the higher branches of criticism was first made explicitly by the German biblical scholar J.G. Eichhorn; the first use of the term "textual criticism" in English dates from the middle of the 19th century. In practice the operations of textual and "higher" criticism cannot be rigidly differentiated: at the very outset of his work a critic, faced with variant forms of a text, inevitably employs stylistic and other criteria belonging to the "higher" branch. The methods of textual criticism, insofar as they are not codified common sense, are the methods of historical inquiry. Texts have been transmitted in an almost limitless variety of ways, and the criteria employed by the textual critic—technical, philological, literary, or aesthetic—are valid only if applied in awareness of the particular set of historical circumstances governing each case.

The value of textual criticism

An acquaintance with the history of texts and the principles of textual criticism is indispensable for the student of history, literature, or philosophy. Written texts supply the main foundation for these disciplines, and some knowledge of the processes of their transmission is necessary for understanding and control of the scholar's basic materials. For the advanced student the criticism and editing of texts offers an unrivalled philological training and a uniquely instructive avenue to the history of scholarship; it is broadly true that all advances in philology have been made in connection with the problems of editing texts. To say this is to recognize that the equipment needed by the critic for his task includes a mastery of the whole field of study within which his text lies; for the editing of Homer (to take an extreme case), a period of some 3,000 years. For the general reader the benefits of textual criticism are less apparent but are nevertheless real. Most men are apt to take texts on trust, even to prefer a familiar version, however debased or unauthentic, to the true one. The reader who resists all change is exemplified by Erasmus' story of the priest who preferred his nonsensical *mumpsimus* to the correct *sumpsimus*. Such people are saved from themselves by the activities of the textual critic.

The law of diminishing returns operates in the textual field as in others: improvements in the texts of the great writers cannot be made indefinitely. Yet a surprisingly large number of texts have not yet been edited satisfactorily. This is particularly true of medieval literature, but also of many modern novels. Indeed the basic materials of most textual investigation, the manuscripts themselves, have as yet not all been identified and catalogued, much less systematically exploited. The first edition of the works of Dickens to be founded on critical study of the textual evidence did not begin to appear until 1966, when K. Tillotson's edition of *Oliver Twist* was published. Reliable principles of Shakespearean editing have begun to emerge only with modern developments in the techniques of analytical bibliography. The Revised Standard Version of the Bible (1952) and the New English Bible (1970) both incorporate readings of the Old Testament unknown before 1947, the year in which early biblical manuscripts—the so-called Dead Sea Scrolls—were discovered in the caves of Qumrān.

The materials of the investigation. The premise of the textual critic's work is that whenever a text is transmitted, variation occurs. This is because human beings are careless, fallible, and occasionally perverse. Variation can occur in several ways: through mechanical damage or accidental omission; through misunderstanding due to changes in fashions of writing; through ignorance of language or subject matter; through inattention or stupidity; and through deliberate efforts at correction. The task of the textual critic is to detect and, so far as possible, undo these effects. His concern is with the reconstruction of what no longer exists. A text is not a concrete artifact, like a pot or a statue, but an abstract concept or idea. The original text of Aeschylus' *Agamemnon* or Horace's *Odes* has perished; what survives is a number of derived forms or states of the text, approximations of varying reliability preserved by tradition. The critic must reduce these approximations as nearly as possible to the first or original state that they imperfectly represent; or if, as sometimes happens for reasons that will be explained below, no single original can be reconstructed or postulated, he must reduce their number to the lowest possible figure. His methods and the degree of his success will be determined by the nature of the individual problem; *i.e.*, the text itself and the circumstances of its transmission. The range of possible situations is vast, as the following survey indicates. The types of text with which the critic is concerned may be classified broadly under three heads.

Books transmitted in print. For practical purposes it is often assumed that the latest edition of a modern book published during the author's lifetime may be treated as the original. This is a simplification. The actual author's original may have been a manuscript or a typescript or a recording; in the process of publication it has passed through several stages of transmission, including possibly storage in a computer, at any one of which errors have necessarily occurred. Experience teaches that some errors will survive uncorrected in the published version. Further errors are likely to occur if a book is reprinted. Even an edition revised by the author is not to be regarded as textually definitive. Errors committed and overlooked by the author himself may be corrected by the critic in appropriate cases. Special problems are posed by an author's second thoughts, whether preserved in his books and papers or incorporated in editions revised by him; recent research has shown that authorial revision in modern printed books has been underestimated. The extent to which a critic is free to choose between authorial variants on aesthetic grounds is a matter of debate.

Books published before the 19th century pose essentially similar problems in a more intractable form, as may be seen in the case of Shakespeare. No manuscript of any of Shakespeare's plays survives, and there were substantial intervals between the dates of composition and the first

printed versions, in which unauthorized variation clearly occurred. For Shakespeare's plays, indeed, the very concept of an author's original may be misleading. Elizabethan printers clearly had little regard for strict textual accuracy, so that allowance must be made not only for error but for deliberate alteration by compositors; thus the textual criticism of 16th- and 17th-century books must include a study of the practices of early printers.

Books transmitted in manuscript. Nearly all classical and patristic texts, and a great many medieval texts, fall into this category. Every handwritten copy of a book is textually unique, and to that extent represents a separate edition of the text. Whereas the characteristic grouping of printed texts is "monogenous" (*i.e.*, in a straight line of descent), that of manuscript texts is "polygenous" or branched and interlocking. The critic is in principle obliged to establish the relationship of every surviving manuscript copy of a text to every other. The difficulty and indeed the feasibility of this undertaking varies enormously from case to case. The following extremes embrace a wide range of intermediate possibilities. (1) The authority for a text may be a single surviving copy (*e.g.*, Menander, *Dyscolus*) or a copy that can be shown to be the source of all other copies (*e.g.*, Varro, *De Lingua Latina*) or an edition printed directly from a copy now lost (*e.g.*, the work of the Roman historian Velleius Paterculus); or a text may be transmitted in scores of copies whose interrelationships cannot be exactly determined (*e.g.*, Claudian, *De raptu Proserpinae*). (2) The interval between the original and the earliest surviving copies may be very short (*e.g.*, the French medieval poet Chrétien de Troyes) or very long (*e.g.*, the Attic tragedians). (3) A tradition may be "dynamic"—*i.e.*, the text may have been copied and recopied many times even in a short time (*e.g.*, Dante's *La divina commedia*); or it may be "static"—*i.e.*, the number of transmissional stages even over a long period may have been few (e.g., *Epigrammata Bobiensia*, a Latin translation of Greek epigrams). (4) A text may be a religious or literary work that was respectfully treated by copyists and protected by an exegetical tradition (*e.g.*, the Bible, the Latin poet Virgil); or a popular book that was exposed to correction, glossing, and amplification by readers (*e.g.*, the *Regula magistri* ["Rule of the Master," a Latin work related to the Rule of St. Benedict] and much medieval vernacular literature). (5) A text may have been written and transmitted after the establishment of a scholarly tradition, or it may show signs of "wild" and arbitrary variation dating from an age in which standards of exact verbal accuracy were low. To this extent all Greek books written before the establishment of the Alexandrian library (see below) were exposed to the hazards associated with oral transmission.

Books transmitted orally. Many texts have been orally transmitted, sometimes for long periods, before being committed to writing, and much textual variation may be attributable to this stage of transmission. Often in such cases the critic cannot attempt to construct an "original" but must stop short at some intermediate stage: thus the edited text of Homer means in practice the closest possible approximation to the text as established by the scholars of Alexandria. The length, complexity, and fidelity of oral traditions varies enormously. The text of the old Indian Rigveda was transmitted orally almost without variation from very ancient to modern times, whereas much old French epic and Provençal lyric has descended in variant redactions for which a common source may be postulated but cannot be reconstructed. Sometimes this is attributable not to spontaneous variation but to deliberate reworking, whether by the author, as appears to be the case with the three versions of the English poem *Piers Plowman*, or by later revisers, as with the four versions of *Digenis Akritas* (a Greek epic). The distinction, however, is not always easy to draw. These considerations apply to a wide range of texts from ancient Hebrew through Old Norse to modern Russian, but they are especially important for medieval literature. In this field perhaps more than in any other the critic's aims and methods will be dictated by the character of the oral tradition, the stage at which it attained a more or less fixed form in writing, and the attitude of copyists in a particular genre to precise verbal accuracy. A problem of particular difficulty and importance is posed by the Greek New Testament. Though the text appears to have been transmitted from the first in writing, the textual variations are in many ways analogous to those of an oral tradition, and it is commonly held that the essential task of the critic is not to try to reconstruct the "original" but to isolate those forms of the text that were current in particular centres in the ancient world.

Critical methods. From the preceding discussion it is apparent that there is only one universally valid principle of textual criticism, the formulation of which can be traced back at least as far as the 18th-century German historian A.L. von Schlözer: that each case is special. The critic must begin by defining the problem presented by his particular material and the consequent limitations of his inquiry. Everything that is said below about "method" must be understood in the light of this general proviso. The celebrated dictum of the 18th-century English classical scholar Richard Bentley that "reason and the facts outweigh a hundred manuscripts" (*ratio et res ipsa centum codicibus potiores sunt*) is not a repudiation of science but a reminder that the critic is by definition one who discriminates (the word itself derives from the Greek word for "judge"), and that no amount of learning or mastery of method will compensate for a lack of common sense. To study the great critics in action is incomparably more instructive than to read theoretical manuals. As the editor of Manilius, A.E. Housman, wrote,

The importance of common sense

> A man who possesses common sense and the use of reason must not expect to learn from treatises or lectures on textual criticism anything that he could not, with leisure and industry, find out for himself. What the lectures and treatises can do for him is to save him time and trouble by presenting to him immediately considerations which would in any case occur to him sooner or later.

Admittedly, the technical advances in textual bibliography mentioned below are not such as would sooner or later occur to any reflective and intelligent person; but bibliography, like paleography, is ancillary to textual criticism proper, and Housman's words are strictly true. What they imply is that good critics are born, not made.

The critical process can be resolved into three stages: (1) recension, (2) examination, and (3) emendation. Though these stages are logically distinct, (2) and (3) are in practice performed simultaneously, and even (1) entails the application of criteria theoretically appropriate to (2) and (3).

The search for the earliest transmitted form of the text

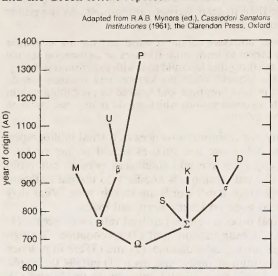
Recension. The operation of recension is the reconstructing of the earliest form or forms of the text that can be inferred from the surviving evidence. Such evidence may be internal or external. Internal evidence consists of all extant copies or editions of the text, together with versions in other languages, citations in other authors, and other sources not belonging to the main textual tradition. These witnesses (as they may be called) must be identified, dated, and described, using the appropriate paleographical and bibliographical techniques. They must then be collated; *i.e.*, the variant readings that they contain must be registered by comparison with some selected form of the text, often a standard printed edition. Where the number of witnesses is large, collation may have to be of selected passages. If there is only one witness to a text, collation and recension are synonymous, and the critic passes straight to examination and emendation. Generally, however, he will be faced with two or more witnesses offering variant forms or states of the text.

Collateral evidence as to the transmission of a text may be supplied from sources external to the direct or indirect textual tradition. Thus the ancient biographers throw light on the circumstances in which Virgil's *Aeneid* was published. Inferred textual stages may be dated on the evidence of copying practices at different periods, or by association with a particular scholar, or from entries in medieval library catalogues. Generally speaking, information of this sort will contribute more to the history than to the criticism of the text, but the two fields are intimately connected; and the better the textual history is known, the more reliable the control of the critic over conjectural solutions to specific problems. In the case of

printed books, such external evidence is as a rule more plentiful; it is often essential, since so much may turn on the accurate dating of editions. Relevant information must be sought in the published and unpublished records of stationers, printers, booksellers, and publishers and in other archival material.

Having assembled his evidence, the critic may proceed, broadly speaking, in one of two different ways, according as he decides to handle the problem of interrelationships "genealogically" or "textually."

In the "genealogical" or "stemmatic" approach, the attempt to reconstruct an original text here relies on the witnesses themselves regarded as physical objects related to each other chronologically and genealogically; the text and the textual vehicle (the book itself) are treated as a single entity. On the basis of shared variants, chiefly errors and omissions, a family tree of the witnesses (*stemma codicum*) is drawn up. Those witnesses that repeat the testimony of other surviving witnesses are discarded, and from the agreements of the remainder the text is reconstructed as it existed in the lost copy from which they descend, the "archetype." Thus in the tradition of the 6th-century monk Cassiodorus' *Institutiones* the relationships of the manuscripts of the authentic version of the text of Book II may be represented by the accompanying diagram. The Roman letters represent extant manuscripts, and the Greek letters represent the lost manuscripts from



Manuscript genealogy, with Greek letters representing sources upon which known manuscripts (Roman letters) must have been based.

which they derive, here arbitrarily dated. The text of the archetype Ω is established by the agreement of B and Σ. Since B survives, the readings of MUP, which are derived from it, would be of value only where B had suffered damage after M and β were copied from it. In such cases the text of β could be inferred from the agreement of UP and the text of B from the agreement of Mβ (or MU or MP). The text of Σ can be inferred from the agreement of SLσ or SL or Sσ (or ST or SD) or Lσ (or LT or LD). K, being copied from L, would be of value only where L had suffered damage after K was copied from it. An important distinction is here exemplified between "trifid" and "bifid" *stemmata*. Where there are three independent witnesses to a source, as with Σ, its reading is certified by the agreement of all three or of any two; where there are only two witnesses, as with Ω, and they disagree, the reading of the source cannot be certified. Even in the latter situation, however, the number of possible variants existing in the source would have been reduced to two. Thus in theory the genealogical, or stemmatic, method allows the critic to eliminate from consideration all variants that cannot be traced back to the archetype or earliest inferable textual state.

While in principle this method is unassailable, it depends for its practical validity on the assumption that each copyist followed only one model or exemplar and generated only variants peculiar to himself. This is called "vertical" transmission, and a tradition of this kind is called "closed." Once the possibility is admitted that a copyist used more than one exemplar or (the more probable supposition) copied an exemplar in which variants from another source or sources had been incorporated—*i.e.*, that more than one textual state may coexist in a single witness—the construction of a *stemma* becomes more complicated and may be impossible. This is called "horizontal" transmission, and a tradition of this kind is called "open" or "contaminated." The practice of critics faced with contamination tends to vary, for historical reasons, from field to field. Editors of classical texts generally adopt a controlled eclecticism, classifying the witnesses broadly by groups according to the general character of their texts and choosing between their readings largely on grounds of intrinsic excellence. Medievalists, following the French scholar Joseph Bédier (see below), sometimes revert to the traditional practice, to which their training may dispose, of selecting a single witness as the main basis of the text. For editors of printed books, contamination is not an important problem.

Analysis of variants

In the "textual" or "distributional" approach, the text and the textual vehicle are dissociated; the emphasis is on the analysis of the variants themselves and their distribution rather than on the character of the text as presented by individual witnesses. The techniques or models employed include those of statistics, symbolic logic, and biological taxonomy. Two theoretical advantages are suggested for this approach. First, objectivity: no judgments of value are entailed, whereas the genealogical method calls for decisions as to the correctness of readings or textual states. Second, the possibility of mechanization: long and elaborate calculations involving thousands of variants may be performed by a computer. This possibility is especially attractive to New Testament critics, who are confronted with about 5,000 manuscripts of the Greek text as well as versions in other languages and patristic citations. In practice, however, these advantages are to a large extent illusory. An "objective" (*i.e.*, undiscriminating) treatment of all variants in a literary text such as Ovid's *Metamorphoses* (of which more than 300 manuscripts exist) without regard to their metrical and stylistic quality would be a self-evident waste of time and produce merely confusion. The critic cannot abrogate his critical function, which implies discrimination, at the very beginning of the critical process. Moreover, the preparation or programming of a text for treatment in this way, whether mechanical aids are used or not, is long and laborious, and one must consider whether in a given case the results justify the expenditure of effort. Texts have been transmitted by a combination of purpose and accident that in any particular instance is both unique and unpredictable, and no machine or statistical model exhibits the versatility necessary to unravel the incomplete and tangled skein. Mechanical methods have been most successful in fields other than recension (see below).

Authenticity of a text

Examination. The process of determining whether the transmitted text or any of the transmitted variants of it is "authentic"—*i.e.*, what the author intended—is known as examination. The prior process of recension has reduced the number of textual states having a claim to be considered "authoritative." Many different situations are possible. In a completely closed tradition it is theoretically feasible to reconstruct the archetype with such certainty that only a single form of the text without variants remains to be examined. In practice this is extremely unlikely to be the situation. Usually the critic is faced with pairs (sometimes triplets) of variants, all with a presumptive claim to be considered authoritative. In some traditions he will confront variant versions of the whole text. Where papyri or other early sources independent of the main tradition are available, he may have to reckon with "pretraditional" (*i.e.*, pre-archetypal) variants. The process of examination calls upon the critic's full range of knowledge as well as his innate powers of taste and discrimination. The criteria applied must be those appropriate to the particular author (supposing his identity to be known), the period, the genre, and the particular character of the work. The opposing demands of analogy and anomaly must be weighed according to the circumstances. Many of the older generation of critics based their decisions on aprioristic or rigidly

analogical principles of elegance and propriety, while the canons of modern criticism are based on historical studies of language and style. It is here that the circularity inherent in the whole operation is most evident, for the linguistic and stylistic criteria employed are themselves based on inductions from texts, probably including the one under examination. There is no escape from this difficulty; as the German philologist Karl Lachmann observed, it is precisely the task of the critic "to tread that circle deftly and warily."

Emendation. The attempt to restore the transmitted text to its authentic state is called emendation. There will usually be a chronological gap, sometimes of several centuries, between the archetype, or earliest inferable state of the text, and the original; nearly all manuscripts of classical authors date from the Middle Ages. The history of the text during the intervening period may be illustrated from external sources; but if examination has convinced the critic that the transmitted text (or its variants) are not authentic, he normally has no recourse but to bridge the gap by conjecture. Conjectural emendation has been defined by the American scholar B.L. Gildersleeve as "the appeal from manuscripts we have to a manuscript that has been lost." Theoretically this definition is acceptable, if we interpret "manuscript" as "source," but in practice the making of conjectures, as distinct from testing them, is intelligent guesswork.

Conjectural emendation

No part of the theory of textual criticism has suffered more from misunderstanding than has conjectural emendation. Such conjectural, or divinatory, criticism has in the past enjoyed a traditional preeminence: Dr. Johnson observed that William Warburton's correction of "good" to "god" in the second act of *Hamlet* (scene 2, line 182) almost set the critic on a level with the author. That idea is as erroneous as the frame of mind in which the Italian scholar C. Pascal founded the Paravia series of editions in order to purge Latin texts of German conjectures. The best critic is he who discriminates best, whether between variants or between transmitted text and conjecture.

Conjectures as a rule occur to the mind spontaneously or not at all; diagnosis and prescription often present themselves at the same moment. This instinctive process is not under the critic's control, though he can sharpen and regulate it by constant study and observation. The outcome of the process, the emendation itself, can and must be controlled and tested by precisely the same criteria as are used in deciding between variants. This is essentially an exercise in balancing probabilities. These probabilities are historical. The conventional distinction between intrinsic and transcriptional (*i.e.*, paleographical or bibliographical) probability tends to obscure a fundamental historical point. If the transmitted form of the text lies at few removes or a short distance in time from the original, a conjectural solution which violates transcriptional probability is less likely to be correct than if the text has undergone a long and complex process of deterioration. In the latter case the critic may attach little or no importance to transcriptional probability. The critic cannot neglect the study of paleography or bibliography, but he must not give them more than their critical due. What that may be depends on the particular historical circumstances. He will study carefully the rationale of error in manuscripts and books themselves rather than in the schematic classifications of critical manuals; and he will learn from experience to distinguish between the types of error that may be called "psychological" (*i.e.*, those committed by a tired or inattentive copyist, whatever language or instruments he uses) and those contingent on the period and the medium of transmission, whether it be the mouth and the ear, the pen, the hand composing stick, the linotype or typewriter keyboard, the computer or photocopying machine, or the printing press. Two complementary principles originated by the New Testament critics of the 18th century are often cited as aids to decision: *utrum in alterum abiturum erat?* ("which reading would be more likely to have given rise to the other?") and *difficilior lectio potior* ("the more difficult reading is to be preferred"). These are no more than useful rules of thumb; it has been suggested that in practice these and other such principles reduce themselves to the truism *melior lectio potior*, "the better reading is to be preferred."

From this discussion it is apparent that the traditional opposition between "conservative" and "radical" styles of criticism that has haunted textual criticism since St. Jerome has no meaning. The critic does not attack or defend the transmitted text; he asks himself whether it is authentic. How radically he treats it, and how many conjectural readings he substitutes for transmitted readings, depends not on his temperament but on the nature of the problem. If he has studied the history of textual criticism he will know that as a matter of demonstrable fact nearly all conjectures are wrong, and he will accept that many of his solutions are in the nature of things provisional.

Editorial technique. Critical texts are edited according to conventions that vary with the type of text (classical, medieval, modern) but follow certain general principles. In some cases, as with newly edited papyri and with palimpsests (writing materials re-used after erasure), the edition will take the form of a diplomatic transcript; *i.e.*, the most accurate possible representation of a particular textual form. Generally, however, the editor constitutes his text in accordance with his own judgment on principles explained in his introduction; and he indicates his sources in critical notes (*apparatus criticus*), preferably at the foot of the page. These notes are usually couched in a special terminology that relies heavily on abbreviation and the use of conventional signs or letters (*sigla*) to identify the witnesses. In classical and patristic texts the language of the notes is usually Latin. Editorial judgment will be influenced by the presumed needs of readers: in an edition intended for scholars, very corrupt passages are often printed as transmitted and marked with a dagger (†), whereas in an edition for the student or general reader some compromise may be accepted in the interests of readability.

A much-discussed problem is the treatment of "accidentals"—variations in spelling, capitalization, punctuation, and the like. Few if any ancient text traditions preserve reliable evidence of authorial practice in these matters, so that the editor is concerned only with variants that affect the sense; in preparing his text for printing he will adopt modern conventions of presentation and punctuation and a normalized orthography. The same holds good for the majority of medieval texts. Printed texts, however, were generally corrected or seen through the press by the author, or at all events by a contemporary, so that the editor may be reasonably confident of reproducing at least a decent approximation to authorial usage. Whether, or to what extent, he should do so is much debated; opinions differ sharply as to the usefulness of "old-spelling" editions of Shakespeare and other early writers.

History of textual criticism. *From antiquity to the Renaissance.* Until the 20th century the development of textual criticism was inevitably dominated by classical and biblical studies. The systematic study and practice of the subject originated in the 3rd century BC with the Greek scholars of Alexandria. Literary culture had before that time been predominantly oral, though books were in common use by the 5th century, and many texts had suffered damage because the idea of precise textual accuracy and reproduction was unfamiliar. The aim of the librarians of Alexandria was to collect and catalogue every extant Greek book and to produce critical editions of the most important together with textual and interpretative commentaries. Many such editions and commentaries did in fact appear. Alexandrian editing was distinguished above all by respect for the tradition; the text was constituted from the oldest and best copies available, and conjectural emendation was rigidly confined to the commentary, which was contained in a separate volume. An elaborate battery of critical signs was used to refer from text to commentary. These techniques were applied, though on a less ambitious scale, by Roman scholars to Latin texts. Fidelity to tradition was the chief legacy of ancient textual scholarship to later ages; the copyist was expected to reproduce his exemplar as exactly as he could, and correction was based on comparison with other copies, not on the unaided conjectural sagacity of the scribe. Such was the practice

The librarians of Alexandria

of the best monastic scriptoria such as that of Tours, or of the best scholars, such as Lupus of Ferrières (fl. 850). From about 1350, however, a change in attitude is evident, particularly in the West. What is often called the revival of learning was in reality a practical movement to enlist the heritage of classical antiquity in the service of the new Christian humanism. In order to make them usable (*i.e.*, readable), texts were corrected freely and often arbitrarily by scholars, copyists, and readers (the three categories being in fact hardly distinguishable). At its best, as seen in the activities of a scholar like Demetrius Triclinius, later medieval and early Renaissance criticism verges on scientific scholarship, but such cases are exceptional. For the most part the correction of texts was a purely subjective display of taste, sometimes right but much more often wrong, and resting as a rule on nothing more solid than a superficial sense of elegance. In consequence, by the 1470s, when the first printed editions (*editiones principes*) of classical texts began to appear, most Greek and Latin authors were circulating in a textually debased condition, and it was manuscripts of this character that almost always served as copy for the early printers. Very little editing in any real sense of the word was done; the scholars who saw the *editiones principes* through the press generally confined themselves to superficial improvements.

Debasement of texts during the Renaissance

From Politian to Cobet. This state of affairs entailed that down to the 19th century most critics were engaged not in establishing and emending texts on scientific principles but in correcting, in a necessarily unsystematic fashion, a vulgate or received text (*lectio recepta*) that was itself the product of an almost entirely haphazard process of variation and conjecture. The situation was aggravated by the fact that the manuscripts themselves, the basic materials of the investigation, were largely inaccessible to scholars. The Italian poet and scholar Politian, unlike most of his contemporaries, was aware that only through the identification and comparison of the best manuscripts could texts be improved; his notes and collations show that he understood the problem correctly as essentially one of control of the sources. What might have been done in this field is shown by his work, cut short by his early death, on the Florentine codex of Justinian's Pandects. Many manuscripts were still privately owned, their very existence unknown to scholars; public libraries were few and published catalogues fewer; travel was difficult, expensive, and often dangerous. It was not until the twin disciplines of diplomatic and paleography were founded by the great Benedictine monks Mabillon and Montfaucon, and developed by their successors, that a critical use of the evidence became possible; and much of the evidence itself did not become available until after the Napoleonic Wars, when most of the private stock of manuscripts passed finally into public collections.

Some advances were taking place, slowly and unsystematically, in both the theory and practice of textual criticism. The history of critical method in this period is most profitably studied in the best editions of the best editors. The accepted method was to correct the text (*i.e.*, the text of the last printed edition) *codicum et ingenii ope*: *i.e.*, with the aid of the manuscript and printed sources and the critic's own ingenuity. Divination was subordinated to authority, and any reading found in a manuscript or printed text was accounted superior to any conjecture, whatever its intrinsic merits. The first important departure from this pattern is seen in the edition of Catullus by J.J. Scaliger (1577), in which the possibilities of the genealogical method, already understood in principle by Politian and other Renaissance scholars, were exemplified by the demonstration that all the extant copies derived from a lost manuscript, whose orthography and provenance Scaliger was prepared to reconstruct. Almost equally significant was Richard Bentley's edition of Horace (1711), in which for the first time the role of conjecture in the critical and editorial process was recognized and the tradition of producing a corrected version of the text of previous editors was decisively rejected. Bentley's scholarship was greatly admired in the Netherlands, and the editions of the great Dutch Latinists J.F. Gronovius and N. Heinsius were informed by Bentleian principles.

Richard Bentley's influence

Under his influence there grew up what may be called an Anglo-Dutch school of criticism, the two most typical representatives of which were Richard Porson and C.G. Cobet. Its strength lay in sound judgment and good taste rooted in minute linguistic and metrical study; its weaknesses were an excessive reliance on analogical criteria and an indifference to German science and method. Its influence may still be seen in the empiricism that characterizes much critical work by English scholars.

From Bentley to Lachmann. The decisive influence on the editing of secular texts came from the New Testament critics of the 18th century. The printed text of the Greek New Testament in common use was still essentially that established in 1516 by Desiderius Erasmus. For his edition, produced in great haste, he had used such manuscripts, neither ancient nor good, as chanced to be accessible to him. Superficially revised, this was the text termed in the Elzevier edition of 1633 "now received by all," *nunc ab omnibus receptum*. Bentley proposed an edition on radical lines in which he engaged to give the text "exactly as it was in the best exemplars at the time of the Council of Nice. So that there shall not be twenty words, nor even particles, difference . . ." This project never materialized, but editions of the Greek text that did not reproduce the *textus receptus* were published in England by Daniel Mace (1729), William Bowyer, the Younger (1763), and Edward Harwood (1776). On the Continent, meanwhile, New Testament criticism was being developed on scientific and historical lines by a succession of distinguished scholars, notably J.A. Bengel, J.J. Wettstein, J.S. Semler, and J.J. Griesbach. They shaped the genealogical method that was later refined by editors of classical texts. Wettstein also deserves commemoration as the first New Testament critic to use *sigla* systematically. This was important, since some at least of the deficiencies of classical editions at this time are attributable to the lack of suitable conventions for the presentation of critical information, together with a conservative and belletristic attitude to technical jargon by publishers, scholars, and users of books in general. Though *sigla* occur sporadically in editions as early as the 16th century and were used by S. Haverkamp in his Lucretius (1725) in something like the modern style, they did not become normal until the second half of the 19th century.

The genealogical, or stemmatic, method of recension has already been described. It is usually associated with the name of the German Karl Lachmann, but it had its origins in the work of J.A. Bengel and his successors, and almost every essential feature of it was already present in the work of Lachmann's precursors such as J.A. Ernesti, F.A. Wolf, K.G. Zumpt, F.W. Ritschl, and J.N. Madvig. Nevertheless Lachmann occupies a central position in the development of textual criticism because of the unusual power and penetration of his scholarship, the range of textual material on which he worked, and his immense contemporary and posthumous influence. His edition of the Greek New Testament (1831; 2nd ed. 1842–50) was intended primarily as a vindication of the principles of Bentley and Bengel and a demonstration that the *textus receptus* must be finally rejected. Similarly his famous edition of Lucretius (1850) is important as an exemplification of the method in action, since the tradition of Lucretius is peculiarly suitable for the purpose. The demonstration fell short of completeness, for Lachmann had not fully grasped the problem and so failed to exploit the method fully. It has been suggested that Lachmann's best critical work was in his editions of medieval German texts; their influence will be considered below. The Lachmannian model of recension derived added authority from seemingly analogous models in other fields, especially that of comparative philology. As propagated by disciples, notably Moritz Haupt, it dominated textual studies for half a century.

The genealogical method of Karl Lachmann

Related developments in the late 19th century. Possibly the most important technical advance in the latter part of the 19th century was the perfection of photography. Instead of travelling in search of his material, the paleographer or critic could now assemble and study it at relatively little expense and without leaving his desk.

During the last quarter of the 19th century the tempo of archaeological discovery in classical and biblical lands

was vastly increased, and many new texts were unearthed. Some of these were in previously unknown languages, setting new problems of decipherment. Specifically relevant to textual studies are the many Greek papyri recovered from Egypt. These have thrown much light on the history and techniques of ancient book production and scholarship and hence, indirectly, on critical problems. Where the texts they contain are already known, their evidence has tended to emphasize our ignorance of the textual history of classical literature in antiquity itself. Being usually far older than the manuscripts already known, they often illuminate the "pretraditional" state of the text; by sometimes offering readings that agree with those of late and "inferior" medieval copies they justify editors in a policy of cautious eclecticism. Papyrus discoveries have been of particular moment for the text of the New Testament.

Editors of printed texts, having invariably received a classical education (no other being available), had naturally followed, with minor modifications, the methods of classical editing. They would reprint the text of the last edition with such improvements as editorial taste and learning suggested but with no attempt to investigate the sources of the text. Since Lachmann's method was inapplicable to printed texts, this procedure continued until, by the end of the 19th century, the text of Shakespeare, for example, was in a state somewhat analogous to that of most classical writers at the time of the *editiones principes.* Much of the work of modern Shakespearean editors has consisted of undoing the damage inflicted by their predecessors. The early 20th century saw the rise of a new school of "biblio-textual" criticism, most notably represented by A.W. Pollard, R.B. McKerrow, and W.W. Greg. Its object was to devise a style of recension appropriate to the special circumstances under which early printed texts were produced and propagated, and its methods were those of analytical bibliography. These developments are of direct importance for the criticism and editing of a large range of texts of the 16th, 17th, and 18th centuries, particularly those of the Elizabethan and Jacobean dramatists. They have also engendered a discussion of general methodological interest on the role of bibliographical as opposed to historical and literary criteria in the editorial process. This debate continues.

Shakespeare texts

Critics and editors of medieval texts had also inevitably been influenced by developments in the classical field. Before Lachmann it had been usual to choose a single manuscript as the main basis for an edition. Because of the circumstances in which much medieval literature was composed and transmitted this was not necessarily unscientific, and the surviving bulk of texts was so large as to dictate that approach in many cases if they were to be edited at all. This had been the style of editing followed by the Belgian Jesuits known as Bollandists, the French Benedictines called Maurists, and the Italian scholar L.A. Muratori, and perpetuated in the indispensable *Patrologiae Cursus Completus* (edition of the Church Fathers) of the French priest Jacques-Paul Migne. At its best it is seen in the editions of medieval Latin chronicles by the 18th-century Oxford antiquary Thomas Hearne, some of which are still standard works. A more scientific approach was adopted in the publications of the *Monumenta Germaniae Historica,* the later volumes of which (from about 1880) were produced by editors trained in the school of Lachmann. Similarly, editors of vernacular texts followed the lead that Lachmann had given in his editions of such early German poems as the *Nibelunge Not* (1826) and the *Iwein* (1827). An important development in the application of the method was due to the medievalists G. Gröber and G. Paris, who first emphasized the significance of common errors. But in the general uncritical enthusiasm for scientific method, the genealogical approach was too often used without regard for the special conditions under which medieval literature has been handed down.

Reaction against the genealogical method. Haupt had proclaimed in his lectures that his main object was to teach method. But confidence in method led to its misuse. The Lachmannian formula of recension was applied to texts, classical as well as medieval, for which it was unsuitable, often with grotesque results. Commonly this took the form of choosing on "scientific" (*i.e.,* stemmatic) grounds a "best manuscript" (*codex optimus*) and defending its readings as authoritative even where common sense showed that they could not be authentic. This was the type of editing satirized by A.E. Housman in the brilliant prefaces to his editions of Manilius (1903) and Juvenal (1905) and in many reviews and articles. It flourished chiefly between 1875 and 1900, but the dangers of excessive methodological rigidity had already been foreseen. In 1841 H. Sauppe in his *Epistola Critica ad G. Hermannum* had emphasized the diversity of transmissional situations and the difficulty or actual impossibility of classifying the manuscripts in all cases. In 1843 Lachmann's pupil O. Jahn, in his edition of Persius, had repudiated the strict application of the genealogical method as unsuitable to the tradition of that poet. The most extreme position was taken by E. Schwartz, who in his edition of Eusebius' *Historia ecclesiastica* (1909) denied that "vertically" transmitted texts of Greek books existed at all. The limitations of the stemmatic method have subsequently been stressed in a more temperate fashion by other writers. The modern tendency is to acknowledge the validity of the method in principle while recommending a cautious empiricism in its application. For the editor of a contaminated tradition—and most traditions are probably contaminated—the lesson of recent research is that authoritative evidence may survive even in late and generally corrupt or interpolated sources.

More radical criticism of the method has come from medievalists. In 1913 and again in 1928 the French scholar J. Bédier attacked the stemmatic method because the *stemmata* it produced for medieval texts almost invariably had only two branches. Subsequent investigation has shown that Bédier overrated the inherent improbability of this situation, and it is generally agreed that his criticisms had to do with improper application rather than with the method itself. The point taken by H. Quentin (1922) has already been mentioned: that the method entails argument in a circle, since it relies on the identification of errors at the beginning of a process designed to lead to that very end. This objection, more cogent in theory than in practice, applies with greater force to medieval than to classical texts. The linguistic and stylistic canons of classical Greek and Latin are relatively strict and well defined, whereas the vocabulary, grammar, and usage of many medieval authors (especially when an oral prehistory is in question) is often not certain enough to allow reliable discrimination between variant and error. Classical texts, moreover, have passed through a series of bottlenecks in their history, which have simplified editorial problems by eliminating a high proportion of the evidence (*cf.* the remarks on papyri above). With a few exceptions, such as the commentary of Servius, only one version of each text remains to be reconstructed, whereas many medieval texts are extant in several redactions that cannot be winnowed by the stemmatic method so as to leave only one. Quentin's own method, which depended on the comparison of variants in groups of three, without prejudice as to their correctness, has not been generally adopted. It is immensely laborious and does not in practice possess the objectivity that its inventor claimed for it. Bédier and Quentin have, however, done good service to textual criticism in enjoining caution. The best critics in all fields now agree in rejecting the "logical" (*i.e.,* the illogical) application of any method if the results conflict with common sense, and in stressing the necessity of judging variant readings and forms of a text on their intrinsic merits in the light of the information available.

Bédier and Quentin

Mechanical methods. Quentin also gave a lead to later investigators in calling attention to the possibility of basing recension on the variants themselves, and the more sophisticated methods of Greg (1927), Archibald Hill (1950), Vinton Dearing (1959), and J. Froger (1968) may be seen as a continuation of his work. It has already been suggested that methods of this type, so far as recension is concerned, have been of primarily theoretical interest. But the use of mechanical and computing techniques in this field is in its infancy, and assessment must be provisional. Certain practical applications seem to have proved them-

selves. Mechanical aids to collation have been successfully used in editing Shakespeare and Dryden. Computer storage and analysis of texts can provide information about authorial usage, such as stylistic and metrical patterns, and facilitate the production of concordances. These aids are more relevant to conjectural emendation (as shown by their application to the Dead Sea Scrolls) and the "higher" criticism (*e.g.*, determination of authenticity) than to the recension of texts. The formula or machine that will do the critic's essential work for him still awaits discovery; the best texts are produced by the best scholars, whatever their method or lack of method. Lachmann observed that the establishment of a text according to its tradition is a strictly historical undertaking. Twentieth-century research into the composition and transmission of ancient, medieval, and modern texts has confirmed the truth of his pronouncement. (E.J.Ke.)

BIBLIOGRAPHY

Historiography: C.V. LANGLOIS and C. SEIGNOBOS, *Introduction aux études historiques* (1898; Eng. trans., *Introduction to the Study of History*, 1898); JAMES HARVEY ROBINSON (ed.), *Selected Essays of J.B. Bury* (1930, reprinted 1964); JAMES T. SHOTWELL, *The History of History* (1939); JAMES W. THOMPSON and BERNARD J. HOLM, *A History of Historical Writing*, 2 vol. (1942); ROBIN G. COLLINGWOOD, *The Idea of History* (1946); HERBERT BUTTERFIELD, *Man on His Past* (1955); JOHN W. MILLER, *The Philosophy of History with Reflections and Aphorisms* (1981); AGNES HELLER, *A Theory of History* (1982).

J.B. BURY, *The Ancient Greek Historians* (1909, reprinted 1958); M.L.W. LAISTNER, *The Greater Roman Historians* (1947, reprinted 1963); MOSES I. FINLEY (ed.), *The Greek Historians* (1959), selected passages in translation with a valuable introduction; MAURICE PLATNAUER (ed.), *Fifty Years of Classical Scholarship* (1954; rev. ed. with appendixes, *Fifty Years (and Twelve) of Classical Scholarship*, 1968), especially chapters 6 by G.T. GRIFFITH and 13 by A.H. MACDONALD; ARNALDO MOMIGLIANO, *Studies in Historiography* (1966), a selection from his vast collection of valuable articles in *Contributo alla storia degli studi classici e del mondo antico*, 5 vol. (1959–69); T.A. DOREY (ed.), *Latin Historians* (1966) and *Latin Biography* (1967); JOHN BARKER, *The Superhistorians: Makers of Our Past* (1982).

GYULA MORAVCSIK, *Byzantinoturcica*, 2nd ed., vol. 1 (1958), not always reliable in its judgments. There is no satisfactory study in English. There is much useful information in the appendixes to J.B. BURY's edition of *The History of the Decline and Fall of the Roman Empire*, 7 vol. (1896–1900, reprinted 1909–14).

THOMAS F. TOUT, "The Study of Mediaeval Chronicles," *Bulletin of the John Rylands Library*, 6:414–438 (1921), reprinted in *The Collected Papers of Thomas Frederick Tout*, 3 vol. (1932–34); REGINALD L. POOLE, *Chronicles and Annals* (1926); M.L.W. LAISTNER, *Thought and Letters in Western Europe, A.D. 500 to 900*, new ed. (1957); CHARLES H. HASKINS, *The Renaissance of the Twelfth Century* (1927); RALPH H.C. DAVIS and JOHN M. WALLACE-HADRILL (eds.), *The Writing of History in the Middle Ages* (1981).

WALLACE K. FERGUSON, *The Renaissance in Historical Thought* (1948); DENYS HAY, "Flavio Biondo and the Middle Ages," *Proceedings of the British Academy*, 45:97–128 (1960); MYRON GILMORE, *Humanists and Jurists* (1963); PAUL O. KRISTELLER, *Eight Philosophers of the Italian Renaissance*, ch. 2 (1964), on Valla; FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965); IDA MAIER, *Ange Politien* (1966); L.D. REYNOLDS and N.G. WILSON, *Scribes and Scholars: A Guide to the Transmission of Greek and Latin Literature* (1968); ROBERTO WEISS, *The Renaissance Discovery of Classical Antiquity* (1969); E.J. KENNEY, "The Character of Humanist Philology," in R.R. BOLGAR (ed.), *Classical Influences on European Culture, A.D. 500–1500* (1971).

J.G.A. POCOCK, *The Ancient Constitution and the Feudal Law: A Study of English Historical Thought in the Seventeenth Century* (1957), also valuable for France; HERBERT BUTTERFIELD, "The History of Historiography and the History of Science," *Mélanges Alexandre Koyré*, vol. 2 (1964); and "Delays and Paradoxes in the Development of Historiography," in KENNETH BOURNE and D.C. WATT (ed.), *Studies in International History: Essays Presented to W. Norton Medlicott* (1967); GLANMOR WILLIAMS, *Reformation Views of Church History* (1970); DONALD R. KELLEY, *Foundations of Modern Historical Scholarship: Language, Law, and History in the French Renaissance* (1970); MAY MCKISACK, *Medieval History in the Tudor Age* (1971); G. STRAUSS, "The Course of German History: The Lutheran Interpretation," in ANTHONY MOLHO and JOHN A. TEDESCHI (eds.), *Renaissance: Studies in Honor of Hans Baron* (1971).

DAVID C. DOUGLAS, *English Scholars, 1660–1730*, 2nd rev. ed. (1951); MARTIN L. CLARKE, *Greek Studies in England, 1700–1830* (1945); DAVID KNOWLES, "Jean Mabillon," *Journal of Ecclesiastical History*, vol. 10, no. 2 (1959), and *Great Historical Enterprises* (1962); CHRISTOPHER DAWSON, "Edward Gibbon," *Proceedings of the British Academy*, 20:159–180 (1934); and EDWARD GIBBON, *The History of the Decline and Fall of the Roman Empire*, 6 vol. (1776–86, best modern edition by J.B. BURY, *op. cit.*); HUGH TREVOR-ROPER, "The Historical Philosophy of the Enlightenment," in *Studies on Voltaire and the Eighteenth Century*, 27:1667–87 (1963), and "The Idea of the Decline and Fall of the Roman Empire," in *The Age of the Enlightenment* (1967).

J.G.D. CLARK, *Prehistoric Europe: The Economic Basis*, rev. ed. (1962); O.G.S. CRAWFORD, *Archaeology in the Field* (1953); J.G.D. CLARK, *World Prehistory* (1960); CHARLES SAMARAN (ed.), *L'Histoire et ses méthodes*, vol. 11 of the *Encyclopédie de la Pléiade* (1961); VIVIAN H. GALBRAITH, *An Introduction to the Study of History* (1964).

ARTHUR M. SCHLESINGER (ed.), *Historical Scholarship in America* (1932); WILLIAM T. HUTCHINSON (ed.), *The Marcus W. Jernegan Essays in American Historiography* (1937); HUGH H. BELLOT, *American History and American Historians* (1952); DONALD SHEEHAN and HAROLD C. SYRETT (eds.), *Essays in American Historiography: Papers Presented in Honor of Allan Nevins* (1960); JOHN HIGHAM (ed.), *The Reconstruction of American History* (1962); THOMAS J. PRESSLY, *Americans Interpret Their Civil War*, 2nd ed. (1962); RICHARD HOFSTADTER, *The Progressive Historians, Turner, Beard, Parrington* (1968); MARCUS CUNLIFFE and ROBIN W. WINKS (eds.), *Pastmasters: Some Essays on American Historians* (1969); C.L. SONNIGSEN, *The Ambidextrous Historian: Historical Writers and Writing in the American West* (1983).

LORD ACTON, *Historical Essays and Studies* (1907), see especially "German Schools of History"; J.B. BURY, *The Idea of Progress: An Inquiry into Its Origin and Growth* (1920, reprinted 1960); EDMUND WILSON, *To the Finland Station* (1940), especially for French historiography; PIETER GEYL, *Napoleon, voor en tegen in de Franse geschiedschrijving* (1946; Eng. trans., *Napoleon. For and Against*, 1949); *Some Modern Historians of Britain: Essays in Honour of R.L. Schuyler* (1951); G.P. GOOCH, *History and Historians in the Nineteenth Century*, 2nd rev. ed. (1952); FERDINAND SCHEVILL, *Six Historians* (1956), particularly interesting on Ranke; GEORG G. IGGERS, "The Image of Ranke in American and German Historical Thought," in *History and Theory*, 2: 17–128 (1962); and *The German Conception of History* (1968); HENRY E. BELL, *Maitland: A Critical Examination and Assessment* (1965); FREDERICK M. BARNARD, *Herder's Social and Political Thought: From Enlightenment to Nationalism* (1965); JOHN CANNON (ed.), *The Historian at Work* (1980); J.W. BURROW, *A Liberal Descent: Victorian Historians and the English Past* (1981).

EUGENE N. ANDERSON, "Meinecke's *Ideengeschichte* and the Crisis in Historical Thinking," in *Medieval and Historiographical Essays in Honor of James Westfall Thompson* (1938); MARC BLOCH, *Apologie pour l'histoire; ou, Métier d'historien* (1949; Eng. trans., *The Historian's Craft*, 1953); *Architects and Craftsmen in History: Festschrift für Abbot Payson Usher* (1956), biographies of leading historians; JACK H. HEXTER, *Reappraisals in History* (1961); RICHARD PARES, *The Historian's Business, and Other Essays* (1961), particularly valuable for the works of ARNOLD TOYNBEE; FRIEDRICH MEINECKE, *Die Idee der Staatsräson in der neueren Geschichte* (1957; Eng. trans., *Machiavellism: The Doctrine of Raison d'État and Its Place in Modern History*, 1957); DONALD C. WATT (ed.), *Contemporary History in Europe* (1968); LAWRENCE STONE, *The Past and the Present* (1981).

On Russian historiography there is no satisfactory general survey in English. The following can be useful for particular periods or historians: ANATOLE G. MAZOUR, *Modern Russian Historiography*, 2nd ed. (1958); ALEXANDER S. VUCINICH, *Science in Russian Culture: A History to 1860* (1963); and RICHARD PIPES (trans.), *Karamzin's Memoir on Ancient and Modern Russia* (1966); JOHN S. CURTISS (ed.), *Essays in Russian and Soviet History, in Honor of Gerold Tanquary Robinson* (1962), especially on Semevsky; ALAN D. FERGUSON and ALFRED LEVIN (eds.), *Essays in Russian History: A Collection Dedicated to George Vernadsky* (1964); JOHN KEEP and LILIANA BRISBY (eds.), *Contemporary History in the Soviet Mirror* (1964).

FRANZ ROSENTHAL, *A History of Muslim Historiography*, 2nd rev. ed. (1968).

WILLIAM G. BEASLEY and E.G. PULLEYBLANK, *Historians of China and Japan* (1961); CHARLES S. GARDNER, *Chinese Traditional Historiography* (1938, reprinted 1961).

Archaeology: Good general introductions to the aims and methods of archaeology are LEONARD WOOLLEY, *Digging Up the Past* (1930); SIR MORTIMER WHEELER, *Archaeology from the Earth* (1954); and GRAHAME CLARK, *Archaeology and Society*, 3rd rev. ed. (1957). For the history of archaeology and its relation to the development of anthropology, see C.W. CERAM,

Götter, Gräber und Gelehrte (1949; Eng. trans., *Gods, Graves and Scholars*, 1951); G. BIBBY, *The Testimony of the Spade* (1956); and GLYN DANIEL, *A Hundred and Fifty Years of Archaeology* (1974). Anthologies of archaeological writings that relate both to the history of the subject and its present methods are many. The following are recommended: R.F. HEIZER, *The Archaeologist at Work* (1959), and *Man's Discovery of his Past*, 2nd ed. (1970); and JACQUETTA HAWKES, *The World of the Past* (1963). For the development of American archaeology, see G. WILLEY and G. SABLOFF, *The History of American Archaeology* (1973). Special aspects of the development of archaeology are dealt with in D. BROTHWELL and E. HIGGS, *Science in Archaeology*, 2nd ed. (1969); GEORGE F. BASS, *Archaeology Under Water* (1967); W.F. LIBBY, *Radiocarbon Dating*, 2nd ed. (1955); KENNETH HUDSON, *A Social History of Archaeology* (1981); MYRA SHACKLEY, *Environmental Archaeology* (1981); and M.G.L. BAILLIE, *Tree-Ring Dating in Archaeology* (1982).

Bibliography: THEODORE BESTERMAN, *The Beginnings of Systematic Bibliography*, 2nd ed. rev. (1936); A. ESDAILE, *A Student's Manual of Bibliography*, 3rd rev. ed. (1954); L.N. MALCLES, *Les Sources du travail bibliographique*, 3 vol. (1950–58); GEORG SCHNEIDER, *Handbuch der Bibliographie*, 4th ed. (1930, reprinted 1969); J.H. SHERA and M.E. EGAN (eds.), *Bibliographic Organization* (1951); RONALD STAVELEY, I.C. MCILWAINE, and JOHN H. ST. MCILWAINE, *Introduction to Subject Study* (1967); R. STOKES, *The Function of Bibliography* (1969); CURT F. BUHLER *et al.*, *Standards of Bibliographical Description* (1949); J.D. COWLEY, *Bibliographical Description and Cataloguing* (1939); R.B. MCKERROW, *An Introduction to Bibliography for Literary Students* (1927); F.T. BOWERS, *Bibliography and Textual Criticism* (1964) and *Principles of Bibliographical Description* (1949); O.M. BRACK, JR. and W. BARNES (eds.), *Bibliography and Textual Criticism: English and American Literature, 1700 to the Present* (1969); ROBERT B. HARMON, *Elements of Bibliography* (1981).

Chronology: JAMES C. MACDONALD, *Chronologies and Calendars* (1897); ALFRED E. STAMP, *Methods of Chronology* (1933); R.L. POOLE, *Studies in Chronology and History*, collected and ed. by A.L. POOLE (1934, reprinted 1969).

On the astronomical basis of Chinese calendrical systems, see JOSEPH NEEDHAM, *Science and Civilisation in China*, vol. 3, pp. 390–408 (1959). A standard reference for conversion between Chinese and Western calendars is MATHIAS TCHANG, *Synchronismes chinois* (1905).

ROBERT SEWELL and S.B. DIKSHITA, *The Indian Calendar* (1896), describes the various systems of calendars in India, with tables of concordance and a useful index; ROBERT SEWELL, *The Siddhantas and the Indian Calendar* (1924), a study of the Hindu astronomical system as a basis for the traditional calendar; SWAMIKANNU PILLAI, *An Indian Ephemeris*, 6 vol. (1922), tables of concordance of Hindu, Muslim, and modern calendars; JEAN FILLIOZAT, "Notions de chronologie," in LOUIS RENOU and JEAN FILLIOZAT, *Inde classique*, vol. 2 (1953), a general summary and list of different eras used in India.

RICHARD A. PARKER, *The Calendars of Ancient Egypt* (1950), a basic work providing valid solutions to most of the problems of Egyptian chronology; "Lunar Dates of Thutmose III and Ramesses II," *J. Near Eastern Stud.*, 16: 39–43 (1957), an analysis of the more important lunar dates of the New Kingdom; M.B. ROWTON, "Manetho's Date for Ramesses II," *J. Egyptian Archaeol.*, 34:57–74 (1948), and "Comparative Chronology at the Time of Dynasty XIX," *J. Near Eastern Stud.*, 19:15–22 (1960), two articles that deal with the date for the accession of Ramses II (1290 or 1304 BC); ERIK HORNUNG, *Untersuchungen zur Chronologie und Geschichte des Neuen Reiches* (1964), the most up-to-date synthesis of the chronological problems of the New Kingdom; I.E.S. EDWARDS, "Absolute Dating from Egyptian Records and Comparison with Carbon-14 Dating," *Phil. Trans. R. Soc.*, Series A, 269:11–18 (1970), includes a complete list of all radiocarbon dates taken from Egyptian samples.

The most complete general work on the chronology of Western Asia, including Mesopotamia, is P.E. VAN DER MEER, *The Chronology of Ancient Western Asia and Egypt*, 2nd rev. ed. (1955); however, it should be used with caution. For the general chronology of ancient Western Asia, with special reference to Mesopotamia, see M.B. ROWTON in *The Cambridge Ancient History*, 3rd ed., vol. 1, pp. 193–237 (1970), and the extensive literature quoted there (up to 1959). For a discussion of the gap in the 2nd millennium BC, see A. GOETZE, "The Kassites and Near Eastern Chronology," *J. Cuneiform Stud.*, 18:97–101 (1964); and B.L. VAN DER WAERDEN, *Die Anfänge der Astronomie*, pp. 28–49 (1966). For the chronology of the late 2nd millennium BC and the chronological gap in the 1st millennium BC, see J.A. BRINKMAN, *A Political History of Post-Kassite Babylonia, 1158–722 B.C.*, pp. 37–85 (1968).

B. RATNER, *Seder Olam Rabba: Die grosse Weltchronik* (1897), the authoritative edition, carefully annotated and with a detailed introduction, of the oldest rabbinic chronology extending from the earliest records to the first century of the current era; E. MAHLER, *Handbuch der jüdischen Chronologie* (1916), the only comprehensive work on Jewish chronology, covering all aspects and summarizing most previous works on the subject; S. ZEITLIN, *Megillat Taanit As a Source for Jewish Chronology and History in the Hellenistic and Roman Periods* (1922), removes all discrepancy in I and II Maccabees' and Josephus' chronological data in the Hasmonean phase of Jewish history, thus investing their statements with historical significance and authority.

ALAN E. SAMUEL, *Greek and Roman Chronology: Calendars and Years in Classical Antiquity* (1972), a fundamental work; HENRY FYNES CLINTON, *Fasti: Hellenici*, 3 vol. (1834), still useful in providing the sources for the framework of Greek chronology; W. KUBITSCHEK, *Grundriss der antiken Zeitrechnung* (1928), the standard handbook but unreliable; A.E. SAMUEL, *Ptolemaic Chronology* (1962); E.J. BICKERMAN, *Chronology of the Ancient World* (1968), a summary survey to be used judiciously; W. DEN BOER, *Laconian Studies*, pt. 1, "The Struggle for the Chronological Pattern" (1954), on ancient attempts to systematize archaic chronology; A.G. WOODHEAD, *The Greeks in the West*, pp. 69–72 (1962), on dates in the archaic period.

E.J. BICKERMAN, *Chronology of the Ancient World* (1968), an excellent short manual, unfortunately marred by a number of factual errors; AGNES KIRSOPP MICHELS, *The Calendar of the Roman Republic* (1967), contains much valuable information on chronology as well as the calendar—probably the best book ever written on the Roman calendar; F.K. GINZEL, *Handbuch der mathematischen und technischen Chronologie das Zeitrechnungswesen der Völker*. 3 vol. (1906–14), the standard work of reference on its subject.

F.K. GINZEL, *Handbuch der mathematischen und technischen Chronologie das Zeitrechnungswesen der Völker*, 3 vol. (1906–14), still the fundamental work; JACK FINEGAN, *Handbook of Bible Chronology* (1964), the best for its subject; E.J. BICKERMAN, *Chronology of the Ancient World* (1968), very thorough and complete.

AL-BIRUNI, *The Chronology of Ancient Nations*, Eng. trans. by C. EDWARD SACHAU (1879), useful source material on the Muslim knowledge of chronology down to the 10th century AD; see especially pp. 16–82. For actual tables of Muslim chronological information and comments, the following may be consulted: E. LACOINE, *Tables de concordance des dates des calendriers . . .* (1891); C.L. IDELER, *Handbuch der mathematischen und technischen chronologie*, 2 vol. (1825–26); S. LANE-POOLE, *The Mohammedan Dynasties* (1894); and C.E. BOSWORTH, *The Islamic Dynasties* (1967). See also articles on calendar, chronology, and "Hidjra" in the *Encyclopaedia of Islam*. An extensive collection of conversion tables is given in FRANK PARISE (ed.), *The Book of Calendars* (1982).

S.G. MORLEY, "An Introduction to the study of the Maya Hieroglyphs," *Bull. U.S. Bur. Am. Ethnol.*, no. 57 (1915, reprinted 1968), a clear exposition of Maya chronology; J. ERIC S. THOMPSON, *Maya Chronology: The Correlation Question* (1935) and *Maya Hieroglyphic Writing: Introduction* (1950, reprinted 1971); H.J. SPINDEN, *The Reduction of Maya Dates* (1924), a presentation of earlier correlation; L. SATTERTHWAITE and E.K. RALPH, "New Radiocarbon Dates and the Maya Correlation Problem," *Am. Antiq.*, 26:165–184 (1960).

Diplomatics: HARRY BRESSLAU, *Handbuch der Urkundenlehre für Deutschland und Italien*, 2nd ed., vol. 1 (1912), vol. 2, pt. 1 (1914), vol. 2, pt. 2, ed. by H.W. KLEWITZ (1931, all reprinted 1958, with separate index), although somewhat out of date, still the best handbook by far for Germany and Italy; OSWALD REDLICH, *Die Privaturkunden des Mittelalters* (1911, reprinted 1967), an excellent work on the private documents of the Middle Ages; LEO SANTIFALLER, *Beiträge zur Geschichte der Beschreibstoffe im Mittelalter*, pt. 1 (1953), the most up-to-date account of the history of writing materials during the Middle Ages; FRANZ DOLGER and JOHANNES KARAYANNOPULOS, *Byzantinische Urkundenlehre die Kaiserurkunden* (1968), the only study of the documents of the Byzantine emperors; C.R. CHENEY, *The Study of the Medieval Papal Chancery* (1966), an excellent general survey of modern research; R.L. POOLE, *Lectures on the History of the Papal Chancery down to the Time of Innocent III* (1915), the best book in English on this subject, though now out of date; PETER HERDE, *Beiträge zum päpstlichen Kanzlei- und Urkundenwesen im 13. Jahrhundert*, 2nd ed. (1967); and *Audientia litterarum contradictarum*, 2 vol. (1970), on papal letters of justice; WILHELM ERBEN, *Die Kaiser- und Königsurkunden des Mittelalters in Deutschland, Frankreich und Italien* (1907, reprinted 1967), an excellent supplement to Bresslau on the imperial and royal documents of the Middle Ages in Germany, France, and Italy; GEORGES TESSIER, *Diplomatique royale française* (1962), the best and most up-to-date handbook on the royal French diplomatic; ALAIN DE BOUARD, *Manuél de diplomatique française et pontificale* (1929), a handbook on French and papal diplomatic;

F.M. STENTON, *The Latin Charters of the Anglo-Saxon Period* (1955), a good brief survey on research since the 18th century; P.H. SAWYER, *Anglo-Saxon Charters* (1968), the most up-to-date annotated list and bibliography; R.C. VAN CAENEGEM, *Royal Writs in England from the Conquest to Glanvill* (1959), the best book on writs; V.H. GALBRAITH, *An Introduction to the Use of the Public Records* (1934) and *Studies in the Public Records* (1948), two works especially useful for the later periods; C.R. CHENEY, *The Records of Medieval England* (1956) and *English Bishops' Chanceries 1100–1250* (1950); T.F. TOUT, *Chapters in the Administrative History of Medieval England,* 6 vol. (1920–33), the basic work on the subject, including the chancery; HARRY BRESSLAU, "International Beziehungen im Urkundenwesen des Mittelalters," *Archiv für Urkundenforschung,* 6:19–76 (1918), the only survey on international relations in the documentary system of the Middle Ages; HORST ENZENSBERGER, *Beiträge zum Kanzlei- und Urkundenwesen der normannischen Herrscher Unteritaliens und Siziliens* (1971), the most up-to-date account of the chancery and documents of the Norman rulers in southern Italy and Sicily; H.O. MEISNER, *Archivalienkunde vom 16. Jahrhundert bis 1918* (1969), the best study of records from the 16th century through 1918 for central Europe, especially Germany; CHARLES CARTER, *The Western European Powers, 1500–1700* (1971), a very useful survey on archives and their records. All the cited works have comprehensive bibliographies, with some of them also containing bibliographies of editions and facsimiles of documents. H. THOMAS HICKERSON, *Archives and Manuscripts: An Introduction to Automated Access* (1981), discusses specific concern of the contemporary archives business.

Epigraphy: J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament,* 3rd ed. (1969), gives translations of inscriptional material, especially from Egypt, Mesopotamia, Palestine, and Asia Minor. J. FRIEDRICH, *Entzifferung verschollener Schriften und Sprachen* (1954; Eng. trans., *Extinct Languages,* 1957), deals especially with the decipherment of inscriptions, while his *Kleinasiatische Sprachdenkmäler* (1932), transliterates a variety of epichoric texts from Asia Minor. R.G. KENT, *Old Persian,* 2nd ed. rev. (1953), is a comprehensive work including grammar, texts, and lexicon. W. HINZ, *Altiranische Funde und Forschungen* (1969), includes a number of important recent inscriptional finds from ancient Iran. J. BLOCH, *Les Inscriptions d'Aśoka* (1950), gives texts and translations of all Aśokan epigraphs; whereas R.B. PANDEY, *Historical and Literary Inscriptions* (1962), provides a transliterated selection of later Indic epigraphy as well. L. DEROY, *Initiation à l'épigraphie mycénienne* (1962), provides a comprehensive survey of the study of Mycenaean tablets. E.S. ROBERTS, *An Introduction to Greek Epigraphy,* 2 vol. (1887–1905), remains the classic compendium of its kind; while more recent, shorter surveys of its subject are found in G. KLAFFENBACH, *Griechische Epigraphik* (1957); and A.G. WOODHEAD, *The Study of Greek Inscriptions* (1959). A more elaborate recent treatment is that of M. GUARDUCCI, *Epigrafia greca,* 2 vol. (1967–69). The Cretan laws have been sumptuously edited and commented in R.F. WILLETTS, "The Law Code of Gortyn," *Kadmos,* suppl. 1 (1967). The old standard works on Latin inscriptions, J.E. SANDYS, *Latin Epigraphy,* 2nd ed. rev. (1927); and R.L.V. CAGNAT, *Cours d'épigraphie latine,* 3rd ed. (1898), are updated only in such summary surveys as R. BLOCH, *L'épigraphie latine,* 3rd ed. (1964). The elaborate edition of the principal Umbrian inscription by J.W. POULTNEY, *The Bronze Tables of Iguvium* (1959), gives text, translation, grammar, extensive commentary, and facsimiles of the tables themselves. R.W.V. ELLIOTT, *Runes: An Introduction* (1959), is especially thorough on the side of British runes; while W. KRAUSE, *Runen* (1970), is a compact but comprehensive survey of the entire field of runology.

Genealogy: For an outline of the sources and methods of procedure, see L.G. PINE, *Heraldry and Genealogy,* 4th ed. (1974). Also useful is A.J. WILLIS, *Genealogy for Beginners,* 3rd. ed. (1976). J. UNETT, *Making a Pedigree,* 2nd ed. (1961), contains information on medieval records. G. HAMILTON EDWARDS, *In Search of Ancestry* (1966), is a larger work with more detail, especially on naval and military sources. On American genealogy, see G.H. DOANE and J.B. BELL, *Searching for Your Ancestors: The How and Why of Genealogy,* 5th ed. (1980); and J.S. SWEET, *Genealogy and Local History: An Archival and Bibliographical Guide,* 2nd rev. ed. (1959). See also L.G. PINE, *American Origins* (1960, reprinted 1980), on sources for genealogical research in Europe, and *The Genealogist's Encyclopedia* (1969). A good starting point in genealogical quests is P. WILLIAM FILBY, *American and British Genealogy and Heraldry: A Selected List of Books,* 3rd ed. (1983).

Paleography: The most comprehensive work, certainly in English, is still E. MAUNDE THOMPSON, *An Introduction to Greek and Latin Palaeography* (1912). Also valuable, and in print, is his shorter version, *Handbook of Greek and Latin Palaeography* (1893, reprinted 1966). For Greek paleography, see B.A. VAN GRONINGEN, *Short Manual of Greek Palaeography,* 3rd rev. ed. (1963); and C.H. ROBERTS, *Greek Literary Hands, 350 B.C.–A.D. 400* (1956). The article "Handwriting" in C.G. CRUMP and E.F. JACOB (eds.), *The Legacy of the Middle Ages* (1926), is an important account of Latin book hands. Essential for the study of abbreviations in medieval Europe is ADRIANO CAPPELLI, *Lexicon abbreviaturarum: dizionario di abbreviature latine ed italiane,* 6th ed. (1961). A standard textbook for medieval English cursive hands is CHARLES JOHNSON and HILARY JENKINSON, *English Court Hand A.D. 1066 to 1500* (1915), continued in Jenkinson's *Later Court Hands in England from the Fifteenth to the Seventeenth Century* (1927). L.C. HECTOR, *The Handwriting of English Documents,* 2nd ed. (1966), contains a valuable introduction to the paleography of English administrative manuscripts. Very many classical, biblical, and liturgical texts have now been published in facsimile, often with colour plates. Large collections in facsimile comprise E.A. LOWE, *Codices latini antiquiores,* 12 vol. (1934–71); CHARLES SAMARAN and ROBERT MARICHAL, *Catalogue des manuscrits en écriture latine, portant des indications de date* (1959–); BERTRAM COLGRAVE (ed.), *Early English Manuscripts in Facsimile* (1951–); and the "Oxford Palaeographical Handbooks," a series designed to deal with various aspects of the subject, such as C.H. ROBERTS (see above); C.E. WRIGHT, *English Vernacular Hands from the Twelfth to the Fifteenth Centuries* (1960); M.B. PARKES, *English Cursive Book Hands, 1250–1500* (1969); and T.A.M. BISHOP, *English Caroline Minuscule* (1971). See also MARC DROGIN, *Medieval Calligraphy, Its History and Technique* (1980); and DONALD JACKSON, *The Story of Writing* (1981).

Sigillography: The best general account is SIR HILARY JENKINSON, *Guide to Seals in the Public Record Office* (1954). JOSEPH H. ROMAN, *Manuel de sigillographie française* (1912), is a longer study mainly on French seals. The basic catalog for European seals and impressions is WALTER DE GRAY BIRCH, *Catalogue of Seals in the Department of Manuscripts in the British Museum,* 6 vol. (1887–1900). Seal matrices are discussed by ALEC B. TONNOCHY in the *Catalogue of British Seal-Dies in the British Museum* (1952). ALFRED B. WYON, *The Great Seals of England from the Earliest Period to the Present Time* (1887), provides a systematic and well-illustrated account of English Great Seals. A full discussion of Classical seal use, with bibliography and occasional reference to Near Eastern sources, is given in the article "Signum," by WENGER in *Pauly-Wissowa Real-Encyclopädie,* vol. 2, col. 2361–2448 (1923); and details of use in the Aegean and Greece are conveniently given in JOHN BOARDMAN, *Greek Gems and Finger Rings* (1970). No adequate study exists of the uses of seals in ancient western Asia. See at present, however, ELENA CASSIN, "Le Sceau: un fait de civilisation dans la Mésopotamie ancienne," *Annales,* pp. 742–751 (1960); and M.A. ROSTOVTSEFF, "Seleucid Babylonia: Bullae and Seals of Clay with Greek Inscriptions," *Yale Classical Studies* 3:1–114 (1932). A very full bibliography of ancient Eastern seal publications from which information on use may be gleaned is given by HANS H. VON DER OSTEN in *Ancient Oriental Seals in the Collection of Mr. Edward T. Newell,* pp. 168–190 (1934); and in *Altorientalische Siegelsteine der Sammlung Hans Silvius von Aulock,* pp. 156–219 (1957), but no more recent convenient bibliography exists; systematic study of use must depend mainly on the examination of sealed texts. For Egypt, P.E. NEWBERRY, *Scarab-Shaped Seals* (1907), provides a brief account, now much in need of revision. Books on Chinese and Japanese seals deal mainly with those of painters, calligraphers, and collectors. ROBERT H. VAN GULIK, *Chinese Pictorial Art As Viewed by the Connoisseur* (1958), has the most illuminating discussion of the artist's use of seals. See also VICTORIA CONTAG and WANG CHI-CH'IEN, *Seals of Chinese Painters and Collectors of the Ming and Ch'ing Periods,* rev. ed. (1966); and CH'EN CHIH-MAI, *Chinese Calligraphers and Their Art* (1966).

Textual criticism: For oral transmission the fundamental work is H.M. and N.K. CHADWICK, *The Growth of Literature,* 3 vol. (1932–40, reprinted 1969). Classical, biblical, and medieval texts are all covered (but somewhat unevenly) by H. HUNGER *et al., Geschichte der Textüberlieferung der antiken und mittelalterlichen Literatur,* 2 vol. (1961–64). For classical texts in general the best introduction is L.D. REYNOLDS and N.G. WILSON, *Scribes and Scholars* (1968); more specialized are A.C. CLARK, *The Descent of Manuscripts* (1918); and A. DAIN, *Les manuscrits,* 2nd rev. ed. (1964). For special areas, see B.A. VAN GRONINGEN, *Traité d'histoire et de critique des textes grecs* (1963); R.D. DAWE, *The Collation and Investigation of Manuscripts of Aeschylus* (1964); R. RENEHAN, *Greek Textual Criticism: A Reader* (1969); L. HAVET, *Manuel de critique verbale appliquée aux textes latins* (1911), an exhaustive catalogue raisonné of scribal error; E.G. TURNER, *Greek Papyri: An Introduction* (1968), important for the history of Greek texts in antiquity. For biblical texts, in addition to Hunger (*op. cit.*), the standard work on the New Testament is B.M. METZGER,

The Text of the New Testament, 2nd ed. (1968). For patristic texts an excellent case study raising important general issues is M. BEVENOT, *The Tradition of Manuscripts: A Study in the Transmission of St. Cyprian's Treatises* (1961). For medieval texts a good exposition of typical problems may be found in D'A.S. AVALLE, "Die Liederhandschriften und die Textkritik," in Hunger (*op. cit.*), 2:273–290; see also J.A. ASHER, "Truth and Fiction: The Text of Medieval Manuscripts," *Aumla*, 25:6–16 (1966). For printed books the standard work is F. BOWERS, *Bibliography and Textual Criticism* (1964); while a useful collection of essays may be found in O.M. BRACK and W. BARNES (eds.), *Bibliography and Textual Criticism: English and American Literature, 1700 to the Present* (1969).

For the history of the stemmatic method and the contribution of Lachmann, S. TIMPANARO, *La genesi del metodo del Lachmann* (1963), is definitive. For its application, P. MAAS, *Textual Criticism* (Eng. trans. 1958), an austere theoretical exposition; and G. PASQUALI, *Storia della tradizione e critica del testo*, 2nd ed. (1952), brilliant but discursive, are complementary and fundamental. The reaction to the stemmatic method may be studied in H. QUENTIN, *Mémoire sur l'établissement du texte de la Vulgate* (1922), *Essais de critique textuelle–ecdotique* (1926); J. BURKE SEVERS, "Quentin's Theory of Textual Criticism," *English Inst. Annual 1941*, pp. 65–93 (1942); J. BEDIER, "La tradition manuscrite du Lai de L'Ombre: Réflexions sur l'art d'éditer les anciens textes," *Romania*, 54:161–96, 321–56 (1928). A. CASTELLANI, *Bédier avait-il raison?* (1957); W.W. GREG, *The Calculus of Variants* (1927); A.A. HILL, "Some Postulates for Distributional Study of Texts," *Stud. in Biblphy.*, 3:63–95 (1950). For further discussion of these developments, see W.P. SHEPARD, "Recent Theories of Textual Criticism," *Mod. Philol.*, 28:129–141 (1930); J. ANDRIEU, "Principes et recherches en critique textuelle," in *Mémorial des études latines . . . J. Marouzeau*, pp. 458–74 (1943); E.B. HAM, "Textual Criticism and Common Sense," *Romance Philol.*, 12:198–215 (1959). For mechanized techniques, see V.A. DEARING, *A Manual of Textual Analysis* (1959); J. FROGER, *La critique des textes et son automatisation* (1968). For the taxonomic approach, see J.G. GRIFFITH, "A Taxonomic Study of the Manuscript Tradition of Juvenal," *Mus. Helv.*, 25:101–138 (1968); "Numerical Taxonomy and Some Primary Manuscripts of the Gospels," *J. Theol. Stud.*, 20:390–406 (1969).

For classical texts on textual criticism the standard works are still: O. STAHLIN, *Editionstechnik*, 2nd ed. (1914); and A. DELATTE and A. SEVERYNS, *Emploi des signes critiques* (1938). For medieval texts, see A. DONDAINE, "Abbréviations latines et signes raccommandés pour l'apparat critique des éditions des textes médiévaux," *Bull. de la soc. int. pour l'étude de la philosophie méd.*, 2:142–149 (1960); and for papyri, see Turner (*op. cit.*); and H.C. YOUTIE, *The Textual Criticism of Documentary Papyri: Prolegomena* (1958).

Many of the works cited above include discussion of general principles of textual criticism. In addition, the following works may be mentioned: H. KANTOROWICZ, *Einführung in die Textkritik* (1921), sound emphasis on the roles of probability and common sense; A.E. HOUSMAN, "The Application of Thought to Textual Criticism," *Proc. Class. Assoc.*, 18:67–84 (1922), a brilliant polemic against hard-and-fast "rules" of criticism; L. BIELER, "The Grammarian's Craft," *Folia*, 10:3–42 (1956), a useful summary discussion of modern developments; and A.H. MCDONALD, "Textual Criticism," in *Oxford Classical Dictionary*, 2nd ed. (1970), a sound exposition for classical students. JEROME J. MCGANN, *A Critique of Modern Textual Criticism* (1983), is a study of the role of non-literary contexts in textual criticism.

(E.B.Fr./G.E.D./F.C.F./W.S.-cg./J.L.A.F./W.H./M.B.R./E.J.Wi./A.G.W./N.A.Z./F.Th./Pe.He./J.Pl./L.G.P./W.G.U./J.Ch./T. C.M./W.W./E.J.Ke./Ed.)

Hitler

The leader of the National Socialist (Nazi) Party and chancellor (*Kanzler*) and Führer ("leader") of Germany, Adolf Hitler was born on April 20, 1889, at Braunau am Inn, Austria. He was chancellor from January 30, 1933, and, after President Paul von Hindenburg's death, assumed the twin titles of Führer and chancellor (August 2, 1934).

Heinrich Hoffmann, Munich



Hitler reviewing troops on the eastern front, 1939.

Hitler's father, Alois (born 1837), was illegitimate. For a time he bore his mother's name, Schicklgruber, but by 1876 he had established his family claim to the surname Hitler. Adolf never used any other surname.

EARLY LIFE AND RISE TO POWER

After his father's retirement from the state customs service, Adolf Hitler spent most of his childhood in Linz, the capital of Upper Austria. It remained his favourite city throughout his life, and he expressed his wish to be buried there. Alois Hitler died in 1903 but left an adequate pension and savings to support his wife and children. Although Hitler feared and disliked his father, he was a devoted son to his mother, who died after much suffering in 1907. With a mixed record as a student, Hitler never advanced beyond a secondary education. After leaving school, he visited Vienna, then returned to Linz, where he dreamed of becoming an artist. Later, he used the small allowance he continued to draw to maintain himself in Vienna. He wished to study art, for which he had some facilities, but he twice failed to secure entry to the Academy of Fine Arts. For some years he lived a lonely and isolated life, earning a precarious livelihood by painting postcards and advertisements and drifting from one municipal hostel to another. Hitler already showed traits that characterized his later life: loneliness and secretiveness, a bohemian mode of everyday existence, and hatred of cosmopolitanism and of the multinational character of Vienna.

In 1913 Hitler moved to Munich. Screened for Austrian military service in February 1914, he was classified as unfit because of inadequate physical vigour; but when World War I broke out he immediately volunteered for the German army and joined the 16th Bavarian Reserve Infantry Regiment. He served throughout the war, was wounded in October 1916, and was gassed two years later. He was hospitalized when the conflict ended. During the war, he was continuously in the front line as a headquarters runner; his bravery in action was rewarded with the Iron Cross, Second Class, in December 1914, and the Iron Cross, First Class (a rare decoration for a corporal), in August 1918. He greeted the war with enthusiasm, as a great relief from the frustration and aimlessness of civilian life. He found discipline and comradeship satisfying and was confirmed in his belief in the heroic virtues of war.

Military service in World War I

Discharged from the hospital amid the social chaos that followed Germany's defeat, Hitler took up political work in Munich in May–June 1919. As an army political agent, he joined the small German Workers' Party in Munich (September 1919). In 1920 he was put in charge of the party's propaganda and left the army to devote himself to improving his position within the party, which in that year was renamed the Nationalsozialistische Deutsche Arbeiterpartei (Nazi). Conditions were ripe for the development of such a party. Resentment at the loss of the war and the severity of the peace terms added to the economic woes and brought widespread discontent. This was especially sharp in Bavaria, due to its traditional separatism and the region's popular dislike of the republican government in Berlin. In March 1920 a coup d'état by a few army officers attempted in vain to establish a right-wing government.

Munich was a gathering place for dissatisfied former servicemen and members of the Freikorps, which had been organized in 1918–19 from units of the German army that were unwilling to return to civilian life, and for political plotters against the republic. Many of these joined the Nazi Party. Foremost among them was Ernst Röhm, a staff member of the district army command, who had joined the German Workers' Party before Hitler and who was of great help in furthering Hitler's rise within the party. It was he who recruited the "strong arm" squads used by Hitler to protect party meetings, to attack socialists and communists, and to exploit violence for the impression of strength it gave. In 1921 these squads were formally organized under Röhm into a private party army, the SA (Sturmabteilung). Röhm was also able to secure protection from the Bavarian government, which depended on the local army command for the maintenance of order and which tacitly accepted some of his terrorist tactics.

Conditions were favourable for the growth of the small party, and Hitler was sufficiently astute to take full advantage of them. When he joined the party, he found it ineffective, committed to a program of nationalist and socialist ideas but uncertain of its aims and divided in its leadership. He accepted its program but regarded it as a means to an end. His propaganda and his personal ambition caused friction with the other leaders of the party. Hitler countered their attempts to curb him by threatening resignation, and because the future of the party depended on his power to organize publicity and to acquire funds, his opponents relented. In July 1921 he became their leader with almost unlimited powers. From the first he set out to create a mass movement, whose mystique and power would be sufficient to bind its members in loyalty to him. He engaged in unrelenting propaganda through the party newspaper, the *Völkischer Beobachter* ("Popular Observer," acquired in 1920), and through meetings whose audiences soon grew from a handful to thousands. With his charismatic personality and dynamic leadership, he attracted a devoted cadre of Nazi leaders, men whose names today live in infamy—Alfred Rosenberg, Rudolf Hess, Hermann Göring, and Julius Streicher.

Leader of the Nazi Party

The Munich *Putsch*

The climax of this rapid growth of the Nazi Party in Bavaria came in an attempt to seize power in the Munich (Beer Hall) *Putsch* of November 1923, when Hitler and General Erich Ludendorff tried to take advantage of the prevailing confusion and opposition to the Weimar Republic to force the leaders of the Bavarian government and the local army commander to proclaim a national revolution. In the melee that resulted, the police and the army fired at the advancing marchers, killing a few of them. Hitler was injured, and four policemen were killed. Placed on trial for treason, he characteristically took advantage of the immense publicity afforded to him. He also drew a vital lesson from the *Putsch*—that the movement must achieve power by legal means. He was sentenced to prison for five years but served only nine months, and those in relative comfort at Landsberg castle. Hitler used the time to dictate the first volume of *Mein Kampf*, his political autobiography as well as a compendium of his multitudinous ideas.

Mein Kampf

Hitler's ideas included inequality among races, nations, and individuals as part of an unchangeable natural order that exalted the "Aryan race" as the creative element of mankind. According to Hitler, the natural unit of mankind was the *Volk* ("the people"), of which the German people was the greatest. Moreover, he believed that the state existed to serve the *Volk*—a mission that to him the Weimar German Republic betrayed. All morality and truth were judged by this criterion: whether it was in accordance with the interest and preservation of the *Volk*. Parliamentary democratic government stood doubly condemned. It assumed the equality of individuals that for Hitler did not exist and supposed that what was in the interests of the *Volk* could be decided by parliamentary procedures. Instead, Hitler argued that the unity of the *Volk* would find its incarnation in the Führer, endowed with perfect authority. Below the Führer the party was drawn from the *Volk* and was in turn its safeguard.

The greatest enemy of Nazism was not, in Hitler's view, liberal democracy in Germany, which was already on the verge of collapse. It was the rival Weltanschauung, Marxism (which for him embraced social democracy as well as communism), with its insistence on internationalism and economic conflict. Beyond Marxism he believed the greatest enemy of all to be the Jew, who was for Hitler the incarnation of evil. There is debate among historians as to when anti-Semitism became Hitler's deepest and strongest conviction. As early as 1919 he wrote, "Rational anti-Semitism must lead to systematic legal opposition. Its final objective must be the removal of the Jews altogether." In *Mein Kampf*, he described the Jew as the "destroyer of culture," "a parasite within the nation," and "a menace."

During Hitler's absence in prison, the Nazi Party languished as the result of internal dissension. After his release, Hitler faced difficulties that had not existed before 1923. Economic stability had been achieved by a currency reform and the Dawes Plan had scaled back Germany's World War I reparations. The republic seemed to have become more respectable. Hitler was forbidden to make speeches, first in Bavaria, then in many other German states (these prohibitions remained in force until 1927–28). Nevertheless, the party grew slowly in numbers, and in 1926 Hitler successfully established his position within it against Gregor Strasser, whose followers were primarily in northern Germany.

Alliance with Alfred Hugenberg

The advent of the Depression in 1929, however, led to a new period of political instability. In 1930 Hitler made an alliance with the Nationalist Alfred Hugenberg in a campaign against the Young Plan, a second renegotiation of Germany's war reparation payments. With the help of Hugenberg's newspapers, Hitler was able for the first time to reach a nationwide audience. The alliance also enabled him to seek support from many of the magnates of business and industry who controlled political funds and were anxious to use them to establish a strong right-wing, anti-socialist government. The subsidies Hitler received from the industrialists placed his party on a secure financial footing and enabled him to make effective his emotional appeal to the lower middle class and the unemployed, based on the proclamation of his faith that Germany would awaken from its sufferings to reassert its natural greatness. Hitler's dealings with Hugenberg and the industrialists exemplify his skill in using those who sought to use him. But his most important achievement was the establishment of a truly national party (with its voters and followers drawn from different classes and religious groups), unique in Germany at the time.

Unremitting propaganda, set against the failure of the government to improve conditions during the Depression, produced a steadily mounting electoral strength for the Nazis. The party became the second largest in the country, rising from 2.6 percent of the vote in the national election of 1928 to more than 18 percent in September 1930. In 1932 Hitler opposed Hindenburg in the presidential election, capturing 36.8 percent of the votes on the second ballot. Finding himself in a strong position by virtue of his unprecedented mass following, he entered into a series of intrigues with conservatives such as Franz von Papen, Otto Meissner, and President Hindenburg's son, Oskar. The fear of communism and the rejection of the Social Democrats bound them together. In spite of a decline in the Nazi Party's votes in November 1932, Hitler insisted that the chancellorship was the only office he would accept. On January 30, 1933, Hindenburg offered him the chancellorship of Germany. His cabinet included few Nazis at that point.

Hitler's personal life had grown more relaxed and stable with the added comfort that accompanied political success. After his release from prison, he often went to live on the Obersalzberg, near Berchtesgaden. His income at this time was derived from party funds and from writing for nationalist newspapers. He was largely indifferent to clothes and food but did not eat meat and gave up drinking beer (and all other alcohols). His rather irregular working schedule prevailed. He usually rose late, sometimes dawdled at his desk, and retired late at night.

At Berchtesgaden, his half sister Angela Raubal and her two daughters accompanied him. Hitler became devoted to one of them, Geli, and it seems that his possessive jealousy drove her to suicide in September 1931. For weeks Hitler was inconsolable. Some time later Eva Braun, a shop assistant from Munich, became his mistress. Hitler rarely allowed her to appear in public with him. He would not consider marriage on the grounds that it would hamper his career. Braun was a simple young woman with few intellectual gifts. Her great virtue in Hitler's eyes was her unquestioning loyalty, and in recognition of this he legally married her at the end of his life.

Eva Braun

DICTATOR, 1933–39

Once in power, Hitler established an absolute dictatorship. He secured the president's assent for new elections. The Reichstag fire, on the night of February 27, 1933 (apparently the work of a Dutch Communist, Marinus van der Lubbe), provided an excuse for a decree overriding all guarantees of freedom and for an intensified campaign of violence. In these conditions, when the elections were held (March 5), the Nazis polled 43.9 percent of the votes. On March 21, the Reichstag assembled in the Potsdam Garrison Church to demonstrate the unity of National Socialism with the old conservative Germany, represented by Hindenburg. Two days later the Enabling Bill, giving full powers to Hitler, was passed in the Reichstag by the combined votes of Nazi, Nationalist, and Centre party deputies (March 23, 1933). Less than three months later all non-Nazi parties, organizations, and labour unions ceased to exist. The disappearance of the Catholic Centre party was followed by a German Concordat with the Vatican in July.

Hitler had no desire to spark a radical revolution. Conservative "ideas" were still necessary if he was to succeed to the presidency and retain the support of the army; moreover, he did not intend to expropriate the leaders of industry, provided they served the interests of the Nazi state. Ernst Röhm, however, was a protagonist of the "continuing revolution"; he was also, as head of the SA, distrusted by the army. Hitler tried first to secure Röhm's support for his policies by persuasion. Hermann Göring and Heinrich Himmler were eager to remove Röhm, but Hitler hesitated until the last moment. Finally, on June 29, 1934, he

reached his decision. On the "Night of the Long Knives," Röhm and his lieutenant Edmund Heines were executed without trial, along with Gregor Strasser, Kurt von Schleicher, and others. The army leaders, satisfied at seeing the SA broken up, approved Hitler's actions. When Hindenburg died on August 2, the army leaders, together with Papen, assented to the merging of the chancellorship and the presidency—with which went the supreme command of the armed forces of the *Reich*. Now officers and men took an oath of allegiance to Hitler personally. Economic recovery and a fast reduction in unemployment (coincident with world recovery, but for which Hitler took credit) made the regime increasingly popular, and a combination of success and police terror brought the support of 90 percent of the voters in a plebiscite.

Hitler devoted little attention to the organization and running of the domestic affairs of the Nazi state. Responsible for the broad lines of policy, as well as for the system of terror that upheld the state, he left detailed administration to his subordinates. Each of these exercised arbitrary power in his own sphere; but by deliberately creating offices and organizations with overlapping authority, Hitler effectively prevented any one of these particular realms from ever becoming sufficiently strong to challenge his own absolute authority.

Foreign policy claimed his greater interest. As he had made clear in *Mein Kampf*, the reunion of the German peoples was his overriding ambition. Beyond that, the natural field of expansion lay eastward, in Poland, the Ukraine, and the U.S.S.R.—expansion that would necessarily involve renewal of Germany's historic conflict with the Slavic peoples, who would be subordinate in the new order to the Teutonic master race. He saw fascist Italy as his natural ally in this crusade. Britain was a possible ally, provided it abandon its traditional policy of maintaining the balance of power in Europe and limit itself to its interests overseas. In the west France remained the natural enemy of Germany and must, therefore, be cowed or subdued to make expansion eastward possible.

Planning eastward expansion

Before such expansion was possible, it was necessary to remove the restrictions placed on Germany at the end of World War I by the Treaty of Versailles. Hitler used all the arts of propaganda to allay the suspicions of the other powers. He posed as the champion of Europe against the scourge of Bolshevism and insisted that he was a man of peace who wished only to remove the inequalities of the Versailles Treaty. He withdrew from the Disarmament Conference and from the League of Nations (October 1933), and he signed a nonaggression treaty with Poland (January 1934). Every repudiation of the treaty was followed by an offer to negotiate a fresh agreement and insistence on the limited nature of Germany's ambitions. Only once did the Nazis overreach themselves: when Austrian Nazis, with the connivance of German organizations, murdered Chancellor Engelbert Dollfuss of Austria and attempted a revolt (July 1934). The attempt failed, and Hitler disclaimed all responsibility. In January 1935 a plebiscite in the Saarland, with a more than 90 percent majority, returned that territory to Germany. In March of the same year, Hitler introduced conscription. Although this action provoked protests from Britain, France, and Italy, the opposition was restrained, and Hitler's peace diplomacy was sufficiently successful to persuade the British to negotiate a naval treaty (June 1935) recognizing Germany's right to a considerable navy. His greatest stroke came in March 1936, when he used the excuse of a pact between France and the Soviet Union to march into the demilitarized Rhineland—a decision that he took against the advice of many generals. Meanwhile the alliance with Italy, foreseen in *Mein Kampf*, rapidly became a reality as a result of the sanctions imposed by Britain and France against Italy during the Ethiopian war. In October 1936, a Rome–Berlin axis was proclaimed by Italian dictator Benito Mussolini; shortly afterward came the Anti-Comintern Pact with Japan; and a year later all three countries joined in a pact. Although on paper France had a number of allies in Europe, while Germany had none, Hitler's Third Reich had become the principal European power.

In November 1937, at a secret meeting of his military leaders, Hitler outlined his plans for future conquest (beginning with Austria and Czechoslovakia). In January 1938 he dispensed with the services of those who were not wholehearted in their acceptance of Nazi dynamism—Hjalmar Schacht, who was concerned with the German economy; Werner von Fritsch, a representative of the caution of professional soldiers; and Konstantin von Neurath, Hindenburg's appointment at the foreign office. In February Hitler invited the Austrian chancellor, Kurt von Schuschnigg, to Berchtesgaden and forced him to sign an agreement including Austrian Nazis within the Vienna government. When Schuschnigg attempted to resist, announcing a plebiscite about Austrian independence, Hitler immediately ordered the invasion of Austria by German troops. The enthusiastic reception that Hitler received convinced him to settle the future of Austria by outright annexation (*Anschluss*). He returned in triumph to Vienna, the scene of his youthful humiliations and hardships. No resistance was encountered from Britain and France. Hitler had taken special care to secure the support of Italy; as this was forthcoming he proclaimed his undying gratitude to Mussolini.

Anschluss

In spite of his assurances that *Anschluss* would not affect Germany's relations with Czechoslovakia, Hitler proceeded at once with his plans against that country. Konrad Henlein, leader of the German minority in Czechoslovakia, was instructed to agitate for impossible demands on the part of the Sudetenland Germans, thereby enabling Hitler to move ahead on the dismemberment of Czechoslovakia. Britain's and France's willingness to accept the cession of the Sudetenland areas to Germany presented Hitler with the choice between substantial gains by peaceful agreement or by a spectacular war against Czechoslovakia. The intervention by Mussolini and British prime minister Neville Chamberlain appear to have been decisive. Hitler accepted the Munich Agreement on September 30. He also declared that these were his last territorial demands in Europe.

Only a few months later, he proceeded to occupy the rest of Czechoslovakia. On March 15, 1939, he marched into Prague declaring that the rest of "Czechia" would become a German protectorate. A few days later (March 23) the Lithuanian government was forced to cede Memel (Klaipeda), next to the northern frontier of East Prussia, to Germany.

Immediately Hitler turned on Poland. Confronted by the Polish nation and its leaders, whose resolution to resist him was strengthened by a guarantee from Britain and France, Hitler confirmed his alliance with Italy (the "Pact of Steel," May 1939). Moreover, on August 23, just within the deadline set for an attack on Poland, he signed a nonagression pact with Joseph Stalin's Soviet Union—the greatest diplomatic bombshell in centuries. Hitler still disclaimed any quarrel with Britain, but to no avail; the German invasion of Poland (September 1) was followed two days later by a British and French declaration of war on Germany.

Invasion of Poland

In his foreign policy, Hitler combined opportunism and clever timing. He showed astonishing skill in judging the mood of the democratic leaders and exploiting their weaknesses—in spite of the fact that he had scarcely set foot outside Austria and Germany and spoke no foreign language. Up to this point every move had been successful. Even his anxiety over British and French entry into the war was dispelled by the rapid success of the campaign in Poland. He could, he thought, rely on his talents during the war as he relied on them before.

WORLD WAR II

Germany's war strategy was assumed by Hitler from the first. When the successful campaign against Poland failed to produce the desired peace accord with Britain, he ordered the army to prepare for an immediate offensive in the west. Bad weather made some of his reluctant generals postpone the western offensive. This in turn led to two major changes in planning. The first was Hitler's order to forestall an eventual British presence in Norway by occupying that country and Denmark in April 1940. Hitler took a close personal interest in this daring operation.

From this time onward his intervention in the detail of military operations grew steadily greater. The second was Hitler's important adoption of General Erich von Manstein's plan for an attack through the Ardennes (which began May 10) instead of farther north. This was a brilliant and startling success. The German armies reached the Channel ports (which they had been unable to reach during World War I) in 10 days. Holland surrendered after 4 days and Belgium after 16 days. Hitler held back General Karl von Rundstedt's tanks south of Dunkirk, thus enabling the British to evacuate most of their army. But the Western campaign as a whole was amazingly successful. On June 10 Italy entered the war on the side of Germany. On June 22 Hitler signed a triumphant armistice with the French on the site of the Armistice of 1918.

Hitler hoped that the British would negotiate an armistice. When this did not happen, he proceeded to plan the invasion of Britain, together with the elimination of British air power. At the same time preparations were begun for the invasion of the Soviet Union, which in Hitler's view was Britain's last hope for a bulwark against German control of the continent. Then Mussolini invaded Greece, where the failures of the Italian armies made it necessary for German forces to come to their aid in the Balkans and North Africa. Hitler's plans were further disrupted by a coup d'état in Yugoslavia in March 1941, overthrowing the government that had made an agreement with Germany. Hitler immediately ordered his armies to subdue Yugoslavia. The campaigns in the Mediterranean theatre, although successful, were limited, compared to the invasion of Russia. Hitler would spare few forces from "Operation Barbarossa," the planned invasion of the Soviet Union.

"Operation Barbarossa"

The attack against the U.S.S.R. was launched on June 22, 1941. The German army advanced swiftly into the Soviet Union, corralling almost three million Russian prisoners, but it failed to destroy its Russian opponent. Hitler became overbearing in his relations with his generals. He disagreed with them about the object of the main attack, and he wasted time and strength by failing to concentrate on a single objective. In December 1941, a few miles before Moscow, a Russian counteroffensive finally made it clear that Hitler's hopes of a single campaign could not be realized.

On December 7, the next day, the Japanese attacked U.S. forces at Pearl Harbor. Hitler's alliance with Japan forced him to declare war on the United States. From this moment on his entire strategy changed. He hoped and tried (like his idol Frederick II the Great) to break what he deemed was the unnatural coalition of his opponents by forcing one or the other of them to make peace. (In the end, the "unnatural" coalition between Stalin and Winston Churchill and Franklin D. Roosevelt did break up, but too late for Hitler.) He also ordered the reorganization of the German economy on a full wartime basis.

Meanwhile, Himmler prepared the ground for a "new order" in Europe. From 1933 to 1939 and in some instances even during the first years of the war, Hitler's purpose was to expel the Jews from the Greater German Reich. In 1941 this policy changed from expulsion to extermination. The concentration camps created under the Nazi regime were thereby expanded to include extermination camps, such as Auschwitz, and mobile extermination squads, the *Einsatzgruppen*. Although Catholics, Poles, homosexuals, Roma (Gypsies), and the handicapped were targeted for persecution, if not outright extermination, the Jews of Germany, Poland, and the Soviet Union were by far the most numerous among the victims; in German-occupied Europe some 6,000,000 Jews were killed during the war. The sufferings of other peoples were only less when measured in their numbers killed.

At the end of 1942, defeat at El-Alamein and at Stalingrad and the American landing in French North Africa brought the turning point in the war, and Hitler's character and way of life began to change. Directing operations from his headquarters in the east, he refused to visit bombed cities or to allow some withdrawals, and he became increasingly dependent on his physician, Theodor Morell, and on the large amounts and varieties of medicines he ingested. Yet Hitler had not lost the power to react vigorously in the face of misfortune. After the arrest of Mussolini in July 1943 and the Italian armistice, he not only directed the occupation of all important positions held by the Italian army but also ordered the rescue of Mussolini, with the intention that he should head a new fascist government. On the eastern front, however, there was less and less possibility of holding up the advance. Relations with his army commanders grew strained, the more so with the growing importance given to the SS (*Schutzstaffel*) divisions. Meanwhile, the general failure of the U-boat campaign and the bombing of Germany made chances of German victory very unlikely.

Assassination attempts

Desperate officers and anti-Nazi civilians became ready to remove Hitler and negotiate a peace. Several attempts on Hitler's life were planned in 1943–44; the most nearly successful was made on July 20, 1944, when Colonel Claus von Stauffenberg exploded a bomb at a conference at Hitler's headquarters in East Prussia. But Hitler escaped with superficial injuries, and, with few exceptions, those implicated in the plot were executed. The reduction of the army's independence was now made complete; National Socialist political officers were appointed to all military headquarters.

Thereafter, Hitler was increasingly ill; but he did not relax or lose control, and he continued to exercise an almost hypnotic power over his close subordinates, none of whom wielded any independent authority. The Allied invasion of Normandy (June 6, 1944) marked the beginning of the end. Within a few months, eight European capitals (Rome, Paris, Brussels, Bucharest, Sofia, Athens, Belgrade, Helsinki) were liberated by the Allies or surrendered to them. In December 1944 Hitler moved his headquarters to the west to direct an offensive in the Ardennes aimed at splitting the American and the British armies. When this failed, his hopes for victory became ever more visionary, based on the use of new weapons (German rockets had been fired on London since June 1944) or on the breakup of the Allied Powers.

After January 1945 Hitler never left the Chancellery in Berlin or its bunker, abandoning a plan to lead a final resistance in the south as the Soviet forces closed in on Berlin. In a state of extreme nervous exhaustion, he at last accepted the inevitability of defeat and thereupon prepared to take his own life, leaving to its fate the country over which he had taken absolute command. Before this, two further acts remained. At midnight on April 28–29 he married Eva Braun. Immediately afterward he dictated his political testament, justifying his career and appointing Admiral Karl Dönitz as head of the state and Josef Goebbels as chancellor.

On April 30 he said farewell to Goebbels and the few others remaining, then retired to his suite and shot himself. His wife took poison. In accordance with his instructions, their bodies were burned.

Hitler's success was due to the susceptibility of postwar Germany to his unique talents as a national leader. His rise to power was not inevitable; yet there was no one who equalled his ability to exploit and shape events to his own ends. The power that he wielded was unprecedented, both in its scope and in the technical resources at its command. His ideas and purposes were accepted in whole or in part by millions of people, especially in Germany but also elsewhere. By the time he was defeated, he had destroyed most of what was left of old Europe, while the German people had to face what they would later call "Year Zero," 1945.

(A.B./W.F.Kn./Ed.)

HITLER'S PLACE IN HISTORY

At the turn of the 21st century more books had been written about Hitler since his death than about Napoleon during the half-century after the latter's demise. Time and distance from the events of World War II have also affected the historical interpretation of Hitler.

There is a general consensus about his historical *importance* (a term that does not imply a positive judgment). Hitler was principally, and alone, responsible for starting World War II. (This was different from the various responsibilities of rulers and of statesmen who had unleashed

Responsibility for the Holocaust

World War I.) His guilt for the implementation of the Holocaust—that is, the shift of German policy from the expulsion to the extermination of Jews, including eventually Jews of all of Europe and of European Russia, is also obvious. Although there exists no single document of his order to that effect, Hitler's speeches, writings, reports of discussions with associates and foreign statesmen, and testimony by those who carried out the actions have often been cited as evidence of his role. Many of his most violent statements were recorded by his minions during his "Table Talks" (including the not entirely authentic "Bormann remarks" of February–April 1945). For example, on January 30, 1939, to celebrate the sixth anniversary of his rule, Hitler told the Reichstag: "Today I will once more be a prophet: If the international Jewish financiers in and outside Europe should succeed in plunging the nations once more in a world war, then the result will not be the Bolshevization of the Earth and thus the victory of Jewry, but the annihilation of the Jewish race in Europe."

In his final will and testament, written just before his suicide in April 1945, he charged the Germans to continue the struggle against the Jews: "Above all, I enjoin the government and the people to uphold the race laws to the limit and to resist mercilessly the poisoner of all nations, international Jewry."

Despite the immense mass of surviving German documents (and the large volume of his recorded speeches and other statements) Hitler was, as he himself said on a few occasions, a secretive man; and some of his views and decisions differed at times from his public expressions.

For a long time historians and other commentators took it for granted that Hitler's wishes and ambitions and ideology were clearly (and frighteningly) set forth in *Mein Kampf*. In the first, autobiographical, portion of *Mein Kampf*, however, he twisted the truth in at least three matters: his relationship to his father (which was very different from the filial affection he had set forth in *Mein Kampf*); the conditions of his life in Vienna (which were less marked by abject poverty than he had stated); and the crystallization of his worldview, including his anti-Semitism, during his Vienna years (the evidence now suggests that this crystallization occurred much later, in Munich).

Mental health

The popular view of Hitler often involves assumptions about his mental health. There has been a tendency to attribute madness to Hitler. Despite the occasional evidences of his furious outbursts, Hitler's cruelties and his most extreme expressions and orders suggest a cold brutality that was fully conscious. The attribution of madness to Hitler would of course absolve him from his responsibility for his deeds and words (as it also absolves the responsibility of those who are unwilling to think further about him). Extensive researches of his medical records also indicate that, at least until the last 10 months of his life, he was not profoundly handicapped by illness (except for advancing symptoms of Parkinson disease). What is indisputable is that Hitler had a certain tendency to hypochondria; that he ingested vast amounts of medications during the war; and that as early as 1938 he convinced himself that he would not live long—which may have been a reason for speeding up his timetable for conquest at that time. It should also be noted that Hitler possessed mental abilities that were denied by some of his earlier critics: these included an astonishing memory for certain details and an instinctive insight into his opponents' weaknesses. Again, these talents increase, rather than diminish, his responsibility for the many brutal and evil actions he ordered and committed.

His most amazing achievement was his uniting the great mass of the German (and Austrian) people behind him. Throughout his career his popularity was larger and deeper than the popularity of the National Socialist Party. A great majority of Germans believed in him until the very end. In this respect he stands out among almost all of the dictators of the 19th and 20th centuries, which is especially impressive when we consider that the Germans were among the best-educated peoples in the 20th century. There is no question that the overwhelming majority of the German people supported Hitler, though often only passively. Their trust in him was greater than their trust in the Nazi hierarchy. Of course, what contributed to this support were the economic and social successes, for which he fully took credit, during his early leadership: the virtual disappearance of unemployment, the rising prosperity of the masses, the new social institutions, and the increase of German prestige in the 1930s—achievements unparalleled in the histories of other modern totalitarian dictatorships. In spite of the spiritual and intellectual progenitors of some of his ideas there is no German national leader to whom he may be compared. In sum, he had no forerunners—another difference between him and other dictators.

By 1938 Hitler had made Germany the most powerful and feared country in Europe (and perhaps in the world). He achieved all of this without war (and there are now some historians who state that had he died in 1938 before the mass executions began, he would have gone down in history as the greatest statesman in the history of the German people). In fact, he came very close to winning the war in 1940; but the resistance of Britain (personified by Winston Churchill) thwarted him. Nevertheless, it took the overwhelming, and in many ways unusual, Anglo-American coalition with the Soviet Union to defeat the Third Reich; and there are reasons to believe that neither side would have been able to conquer him alone. At the same time it was his brutality and some of his decisions that led to his destruction, binding the unusual alliance of capitalists and communists, of Churchill and Roosevelt and Stalin together. Hitler thought he was a great statesman, but he did not realize the unconditional contemptibility of what he had unleashed; he thought that the coalition of his enemies would eventually break up, and then he would be able to settle with one side or the other. In thinking thus he deceived himself, though such wishes and hopes were also current among many Germans until the end.

Open and hidden admirers of Hitler continue to exist (and not only in Germany): some of them because of a malign attraction to the efficacy of evil; others because of their admiration of Hitler's achievements, no matter how transitory or brutal. However, because of the brutalities and the very crimes associated with his name, it is not likely that Hitler's reputation as the incarnation of evil will ever change. (J.Lu.)

BIBLIOGRAPHY

Writings and speeches: Hitler's speeches have been collected by MAX DOMARUS, *Hitler: Speeches and Proclamations, 1932–1945* (1990– ; originally published in German, 2 vol., 1962). Hitler's words are also recorded in *Secret Conversations, 1941–1944* (1953, reissued as *Hitler's Secret Conversations, 1941–1944*, 1976; also published as *Hitler's Table Talk, 1941–44: His Private Conversations*, 2nd ed., 1973). The first part of *Mein Kampf*, trans. from German by JAMES MURPHY, 2 vol. in 1 (1939, reissued 1981; originally published in German, 2 vol., 1925–27), is his autobiography. *Hitler's Secret Book* (1961, reprinted 1986), is a translation of a manuscript dictated by Hitler in 1928.

Biographies: Biographical studies include IAN KERSHAW, *Hitler, 1889–1936: Hubris* (1998, reissued 2000); ALAN BULLOCK, *Hitler: A Study in Tyranny*, completely rev. ed. (1962, reissued 1995), also available in an abridged ed. with the same title (1971, reissued 1991); JOACHIM C. FEST, *Hitler* (1974, reissued 1992; originally published in German, 1973); BRADLEY F. SMITH, *Adolf Hitler: His Family, Childhood, and Youth* (1967, reissued 1979); WILLIAM CARR, *Hitler: A Study in Personality and Politics* (1978, reprinted 1986); CHARLES BRACELEN FLOOD, *Hitler: The Path to Power* (1989), which traces Hitler's life and politics through 1923; KONRAD HEIDEN, *Der Fuehrer: Hitler's Rise to Power*, trans. from German by RALPH MANHEIM (1944, reissued 1969; also published as *The Führer*, 1999), which deals with the period up to 1934; HUGH TREVOR-ROPER, *The Last Days of Hitler*, 7th ed. (1995); and SEBASTIAN HAFFNER, *The Meaning of Hitler* (1979, reissued 1997; originally published in German, 1978), a profound and well-written biographical essay. EBERHARD JÄCKEL, *Hitler in History* (1984), briefly examines Hitler's rise to power and military involvement; and JOHN LUKACS, *The Hitler of History* (1997), is a study of the biography and biographers of Hitler.

Reminiscences: Reminiscences of Hitler include OTTO WAGENER, *Hitler: Memoirs of a Confidant*, ed. by HENRY ASHBY TURNER, JR. (1985, reissued 1987; originally published in German, 1978), recounting the memories of a Nazi Party official; and GERTRAUD JUNGE, *Voices from the Bunker*, ed. by PIERRE GALANTE and EUGÈNE SILIANOFF (also published as *Last Witnesses in the Bunker*, 1989; originally published in French, 1989),

translating the memoirs of Hitler's private secretary from 1943 to 1945.

Special topics. Other topics are dealt with in HAROLD J. GORDON, JR., *Hitler and the Beer Hall Putsch* (1972); ERNST HANFSTAENGL, *Unheard Witness* (1957; also published as *Hitler: The Missing Years*, 1957, reissued 1994), covering the years 1922–34; ALBERT SPEER, *Inside the Third Reich* (1970, reissued 1997; originally published in German, 1969); and IAN KERSHAW, *The "Hitler Myth": Image and Reality in the Third Reich* (1987, reissued 1989; originally published in German, 1980), discussing Hitler's image as portrayed through German propaganda. RICHARD F. HAMILTON, *Who Voted for Hitler?* (1982); and THOMAS CHILDERS, *The Nazi Voter* (1983), analyze his political support. GERALD FLEMING, *Hitler and the Final Solution* (1984, reissued 1994 with new documentation; originally published in German, 1982), reviews Hitler's connection with the mass exterminations. DANIEL JONAH GOLDHAGEN, *Hitler's Willing Executioners: Ordinary Germans and the Holocaust* (1996), is a controversial work exploring the rise of anti-Semitism in Germany and the complicity of ordinary Germans in the Holocaust. Two important recent contributions are BRIGITTE HAMANN, *Hitler's Vienna: A Dictator's Apprenticeship* (1999; originally published in German, 1996); and FRITZ REDLICH, *Hitler: Diagnosis of a Destructive Prophet* (1999), a medical and psychological history.

(A.B./W.F.K.n./J.Lu./Ed.)

The Holocaust

The Holocaust (Hebrew: Sho'ah; Yiddish and Hebrew: Ḥurban ["Destruction"]) was the systematic state-sponsored killing of six million Jewish men, women, and children and millions of others by Nazi Germany and its collaborators during World War II. The Germans called this "the final solution to the Jewish question." The word *Holocaust* is derived from the Greek *holokauston*, a translation of the Hebrew word *'olah*, meaning a burnt sacrifice offered whole to God. This word was chosen because in the ultimate manifestation of the Nazi killing program—the extermination camps—the bodies of the victims were consumed whole in crematoria and open fires.

Prelude to the Holocaust 629
Nazi anti-Semitism
Kristallnacht
Non-Jewish victims of Nazism
Nazi expansion and the formation of ghettos
Systematic killing 630
The *Einsatzgruppen*
The extermination camps
Jewish resistance
Concealment and liberation
The aftermath 632
Displacement of the survivors
Trial of the perpetrators
Legacy of the Holocaust
Artistic responses to the Holocaust
Conclusion 633
Bibliography 633

PRELUDE TO THE HOLOCAUST

Nazi anti-Semitism. Even before the Nazis came to power in Germany in 1933, they had made no secret of their anti-Semitism. As early as 1919, Adolf Hitler had written, "Rational anti-Semitism, however, must lead to systematic legal opposition. . . . Its final objective must unswervingly be the removal of the Jews altogether." In *Mein Kampf* ("My Struggle"; 1925–27), Hitler further developed the idea of the Jews as an evil race struggling for world domination. Nazi anti-Semitism was rooted in religious anti-Semitism and enhanced by political anti-Semitism. To this the Nazis added a further dimension: racial anti-Semitism. Nazi racial ideology characterized the Jews as *Untermenschen* (German: "subhumans"). The Nazis portrayed Jews as a race and not a religious group. Religious anti-Semitism could be resolved by conversion, political anti-Semitism by expulsion. Ultimately, the logic of Nazi racial anti-Semitism led to annihilation.

When Hitler came to power legally on January 30, 1933, as the head of a coalition government, his first objective was to consolidate power and to eliminate political opposition. The assault against the Jews began on April 1 with a boycott of Jewish businesses. A week later the Nazis dismissed Jews from the civil service, and by the end of the month, the participation of Jews in German schools was restricted by a quota. On May 10, thousands of Nazi students, together with many professors, stormed university libraries and bookstores in 30 cities throughout Germany to remove tens of thousands of books written by non-Aryans and those opposed to Nazi ideology. The books were tossed into bonfires in an effort to cleanse German culture of "un-Germanic" writings. A century earlier, Heinrich Heine—a German poet of Jewish origin—had said, "Where one burns books, one will, in the end, burn people." In Nazi Germany, the time between the burning of Jewish books and the burning of Jews was eight years.

As discrimination against Jews increased, German law required a legal definition of a Jew and an Aryan. Promulgated at the annual Nazi Party rally in Nürnberg on September 15, 1935, the Nürnberg Laws—the Law for the Protection of German Blood and German Honour and the Law of the Reich Citizen—became the centerpiece of anti-Jewish legislation and a precedent for defining and categorizing Jews in all German-controlled lands. Marriage and sexual relations between Jews and citizens of "German or kindred blood" were prohibited. Only "racial" Germans were entitled to civil and political rights. Jews were reduced to subjects of the state. The Nürnberg Laws formally divided Germans and Jews, yet neither the word *German* nor the word *Jew* was defined. That task was left to the bureaucracy. Two basic categories were established in November: Jews—those with at least three Jewish grandparents—and *Mischlinge* ("mongrels," or "mixed breeds")—people with one or two Jewish grandparents. Thus, the definition of a Jew was primarily based not on the identity an individual affirmed or the religion he practiced but on his ancestry. Categorization was the first stage of destruction. Nürnberg Laws

Responding with alarm to Hitler's rise, the Jewish community sought to defend their rights as Germans. For those Jews who felt themselves fully German and who had patriotically fought in World War I, the Nazification of German society was especially painful. Zionist activity intensified. "Wear it with pride," journalist Robert Wildest wrote in 1933 of the Jewish identity the Nazis had so stigmatized. Martin Buber led an effort at Jewish adult education, preparing the community for the long journey ahead. Rabbi Leo Baeck circulated a prayer for Yom Kippur (the Day of Atonement) in 1935 that instructed Jews how to behave: "We bow down before God; we stand erect before man." Yet while few, if any, could foresee its eventual outcome, the Jewish condition was increasingly perilous and expected to get worse.

By the late 1930s there was a desperate search for countries of refuge. Those who could get visas and qualify under stringent quotas emigrated to the United States. Many went to Palestine, where the small Jewish community was willing to receive refugees. Still others sought refuge in neighbouring European countries. Most countries, however, were unwilling to receive large numbers of refugees. Refuge

Responding to domestic pressures to act on behalf of Jewish refugees, U.S. President Franklin D. Roosevelt convened, but did not attend, the Évian Conference on

resettlement, in Évian-les-Bains, France, in July 1938. In his invitation to government leaders, Roosevelt specified that they would not have to change laws or spend government funds; only philanthropic funds would be used for resettlement. The result was that little was attempted, and less accomplished.

Kristallnacht. On the evening of November 9, 1938, carefully orchestrated anti-Jewish violence "erupted" throughout the Reich, which since March had included Austria. Over the next 48 hours rioters burned or damaged more than 1,000 synagogues and ransacked and broke the windows of more than 7,500 businesses. The Nazis arrested some 30,000 Jewish men between the ages of 16 and 60 and sent them to concentration camps. Police stood by as the violence—often the action of neighbours, not strangers—occurred. Firemen were present not to protect the synagogues but to ensure that the flames did not spread to adjacent "Aryan" property. The pogrom was given a quaint name: *Kristallnacht* ("Crystal Night," or "Night of Broken Glass"). In its aftermath, Jews lost the illusion that they had a future in Germany.

On November 12, 1938, Field Marshall Hermann Göring convened a meeting of Nazi officials to discuss the damage to the German economy from pogroms. The Jewish community was fined one billion Reichsmarks. Moreover, Jews were made responsible for cleaning up the damage. German Jews, but not foreign Jews, were barred from collecting insurance. In addition, Jews were soon denied entry to theatres, forced to travel in separate compartments on trains, and excluded from German schools. These new restrictions were added to earlier prohibitions, such as those barring Jews from earning university degrees, from owning businesses, or from practicing law or medicine in the service of non-Jews. The Nazis would continue to confiscate Jewish property in a program called "Aryanization." Göring concluded the November meeting with a note of irony: "I would not like to be a Jew in Germany!"

Non-Jewish victims of Nazism. While Jews were the primary victims of Nazism as it evolved and were central to Nazi racial ideology, other groups were victimized as well—some for what they did, some for what they refused to do, and some for what they were.

Political dissidents, trade unionists, and Social Democrats were among the first to be arrested and incarcerated in concentration camps. Under the Weimar government, centuries-old prohibitions against homosexuality had been overlooked, but this tolerance ended violently when the SA (Storm Troopers) began raiding gay bars in 1933. Homosexual intent became just cause for prosecution. The Nazis arrested German and Austrian male homosexuals—there was no systematic persecution of lesbians—and interned them in concentration camps, where they were forced to wear special yellow armbands and later pink triangles. Jehovah's Witnesses were a problem for the Nazis because they refused to swear allegiance to the state, register for the draft, or utter the words "Heil Hitler." As a result the Nazis imprisoned many of the roughly 20,000 Witnesses in Germany. The Nazis also singled out the Roma (Gypsies). They were the only other group that the Nazis systematically killed in gas chambers alongside the Jews.

Persecution of homosexuals

In 1939 the Germans initiated the T4 Program—framed euphemistically as a "euthanasia" program—for the murder of mentally retarded, physically disabled, and emotionally disturbed Germans who departed from the Nazi ideal of Aryan supremacy. The Nazis pioneered the use of gas chambers and mass crematoria under this program.

Following the invasion of Poland, German occupation policy especially targeted the Jews but also brutalized non-Jewish Poles. In pursuit of *Lebensraum* ("living space"), Germany sought systematically to destroy Polish society and nationhood. The Nazis killed Polish priests and politicians, decimated the Polish leadership, and kidnapped the children of the Polish elite, who were raised as "voluntary Aryans" by their new German "parents." Many Poles were also forced to perform hard labour on survival diets, deprived of property and uprooted, and interned in concentration camps.

Nazi expansion and the formation of ghettos. Paradoxically, at the same time that Germany tried to rid itself of its Jews via forced emigration, its territorial expansions kept bringing more Jews under its control. Germany annexed Austria in March 1938 and the Sudetenland (now in the Czech Republic) in September 1938. It established control over the Protectorate of Bohemia and Moravia (now in the Czech Republic) in March 1939. When Germany invaded Poland on September 1, 1939, the "Jewish question" became urgent. When the division of Poland between Germany and the Soviet Union was complete, more than two million more Jews had come under German control. For a time, the Nazis considered shipping the Jews to the island of Madagascar, off the southeast coast of Africa. But as the seas became a war zone and the resources required for such a massive deportation scarce, they discarded the plan as impractical.

The Judenräte

On September 21, 1939, Reinhard Heydrich ordered the establishment of the Judenräte ("Jewish Councils"), comprising up to 24 men—rabbis and Jewish leaders. Heydrich's order made these councils personally responsible in "the literal sense of the term" for carrying out German orders. When the Nazis sealed the Warsaw Ghetto, the largest of German-occupied Poland's 400 ghettos, in the fall of 1940, the Jews—then 30 percent of Warsaw's population—were forced into 2.4 percent of the city's area. The ghetto's population reached a density of over 200,000 persons per square mile (77,000 per square km) and 9.2 per room. Disease, malnutrition, hunger, and poverty took their toll even before the first bullet was fired.

For the German rulers, the ghetto was a temporary measure, a holding pen for the Jewish population until a policy on its fate could be established and implemented. For the Jews, ghetto life was the situation under which they thought they would be forced to live until the end of the war. They aimed to make life bearable, even under the most trying circumstances. When the Nazis prohibited schools, they opened clandestine schools. When the Nazis banned religious life, it persisted in hiding. The Jews used humour as a means of defiance, so too song. They resorted to arms only late in the Nazi assault.

Historians differ on the date of the decision to murder Jews systematically, the so-called "final solution to the Jewish question." There is debate about whether there was one central decision or a series of regional decisions in response to local conditions; but in either case, when Germany attacked the Soviet Union, its former ally, in June of 1941, the Nazis began the systematic killing of Jews.

SYSTEMATIC KILLING

The *Einsatzgruppen*. Entering conquered Soviet territories alongside the Wehrmacht (the German armed forces) were 3,000 men of the *Einsatzgruppen* ("deployment groups"), special mobile killing units. Their task was to murder Jews, Soviet commissars, and Roma in the areas conquered by the army. Alone or with the help of local police, native anti-Semitic populations, and accompanying Axis troops, the *Einsatzgruppen* would enter a town, round up their victims, herd them to the outskirts of the town, and shoot them. They killed Jews in family units. Just outside Kiev, Ukraine, in the valley of Baby Yar, an *Einsatzgruppe* killed 33,771 Jews on September 28–29, 1941. In the Rumbula Forest outside the ghetto in Riga, Latvia, 25,000–28,000 Jews died on November 30 and December 8–9. Beginning in the summer of 1941, *Einsatzgruppen* killed more than 70,000 Jews at Ponary, outside Vilna (now Vilnius) in Lithuania. They slaughtered 9,000 Jews, half of them children, at the Ninth Fort adjacent to Kovno (now Kaunas), Lithuania, on October 28.

The mass shootings continued unabated, with a first wave and then a second. When the killing ended in the face of a Soviet counteroffensive, special units returned to dig up the dead and burn their bodies to destroy the evidence of the crimes. It is estimated that the *Einsatzgruppen* killed more than one million people, most of whom were Jews.

Historians are divided about the motivations of the members of *Einsatzgruppen*. Christopher Browning describes them as ordinary men in extraordinary circumstances in which conformity, peer pressure, careerism, obedience to orders, and group solidarity gradually overcame moral inhibitions. Daniel Goldhagen sees them as

"willing executioners," sharing Hitler's vision of genocidal anti-Semitism and finding their tasks unpleasant but necessary. Both concur that no *Einsatzgruppe* member faced punishment if he asked to be excused. Individuals had a choice whether to participate or not. Almost all chose to become killers.

The extermination camps. On January 20, 1942, Reinhard Heydrich convened the Wannsee Conference at a lakeside villa in a Berlin suburb to organize the "final solution to the Jewish question." Around the table were 15 men representing government agencies necessary to implement so bold and sweeping a policy. The language of the meeting was clear, but the meeting notes were circumspect: "Another possible solution to the [Jewish] question has now taken the place of emigration, *i.e.,* evacuation to the east. . . . Practical experience is already being collected which is of the greatest importance in the relation to the future final solution of the Jewish question." Participants understood "evacuation to the east" to mean deportation to killing centres.

The "final solution"

In early 1942 the Nazis built extermination camps at Treblinka, Sobibor, and Belzec in Poland. The death camps were to be the essential instrument of the "final solution." The *Einsatzgruppen* had traveled to kill their victims. With the extermination camps, the process was reversed. The victims traveled by train, often in cattle cars, to their killers. The extermination camps became factories producing corpses, effectively and efficiently, at minimal physical and psychological cost to German personnel. Assisted by Ukrainian and Latvian collaborators and prisoners of war, a few Germans could kill tens of thousands of prisoners each month. At Chelmno, the first of the extermination camps, the Nazis used mobile gas vans. Elsewhere, they built permanent gas chambers linked to the crematoria where bodies were burned. Carbon monoxide was the gas of choice at most camps. Zyklon-B, an especially lethal killing agent, was employed primarily at Auschwitz and later at other camps.

Auschwitz

Auschwitz, perhaps the most notorious and lethal of the concentration camps, was actually three camps in one: a prison camp (Auschwitz I), an extermination camp (Auschwitz II–Birkenau), and a slave-labour camp (Auschwitz III–Buna-Monowitz). Upon arrival, Jewish prisoners faced what was called a *Selektion.* A German doctor presided over the selection of pregnant women, young children, the elderly, handicapped, sick, and infirm for immediate death in the gas chambers. As necessary, the Germans selected able-bodied prisoners for forced labour in the factories adjacent to Auschwitz where one German company, IG Farben, invested 700,000 million Reichsmarks in 1942 alone to take advantage of forced labour. Deprived of adequate food, shelter, clothing, and medical care, these prisoners were literally worked to death. Periodically, they would face another *Selektion.* The Nazis would transfer those unable to work to the gas chambers of Birkenau.

While the death camps at Auschwitz and Majdanek used inmates for slave labour to support the German war effort, the extermination camps at Belzec, Treblinka, and Sobibor had one task alone: killing. At Treblinka, a staff of 120, of whom only 30 were SS (the Nazi paramilitary corps), killed some 750,000 to 900,000 Jews during the camp's 17 months of operation. At Belzec, German records detail a staff of 104, including about 20 SS, who killed some 600,000 Jews in less than 10 months. At Sobibor, they murdered about 250,000. These camps began operation during the spring and summer of 1942, when the ghettos of German-occupied Poland were filled with Jews. Once they had completed their missions—murder by gassing, or "resettlement in the east," to use the language of the Wannsee protocols—the Nazis closed the camps. There were six extermination camps, all in German-occupied Poland, among the thousands of concentration and slave-labour camps throughout German-occupied Europe.

Holocaust in Hungary

The impact of the Holocaust varied from region to region, and from year to year in the 21 countries that were directly affected. Nowhere was the Holocaust more intense and sudden than in Hungary. What took place over several years in Germany occurred over 16 weeks in Hungary. Entering the war as a German ally, Hungary had persecuted its Jews but not permitted their deportation. After Germany invaded Hungary on March 19, 1944, this situation changed dramatically. By mid-April the Nazis had confined Jews to ghettos. On May 15, deportations began, and over the next 55 days, the Nazis deported some 438,000 Jews from Hungary to Auschwitz on 147 trains.

Policies differed widely among Germany's Balkan allies. In Romania it was primarily the Romanians themselves who slaughtered the country's Jews. Toward the end of the war, however, when the defeat of Germany was all but certain, the Romanian government found more value in living Jews who could be held for ransom or used as leverage with the West. Bulgaria permitted the deportation of Jews from neighbouring Thrace and Macedonia, but government leaders faced stiff opposition to the deportation of native Bulgarian Jews.

German-occupied Denmark rescued most of its own Jews by spiriting them to Sweden by sea in October 1943. This was possible partly because the German presence in Denmark was relatively small. Moreover, while anti-Semitism in the general population of many other countries led to collaboration with the Germans, Jews were an integrated part of Danish culture. Under these unique circumstances, Danish humanitarianism flourished.

In France, Jews under Fascist Italian occupation in the southeast fared better than the Jews of Vichy France, where collaborationist French authorities and police provided essential support to the understaffed German forces. The Jews in those parts of France under direct German occupation fared the worst. Although allied with Germany, the Italians did not participate in the Holocaust until Germany occupied northern Italy after the overthrow of the Fascist leader, Benito Mussolini.

Throughout German-occupied territory the situation of Jews was desperate. They had meagre resources and few allies and faced impossible choices. A few people came to their rescue, often at the risk of their own lives. Swedish diplomat Raoul Wallenberg arrived in Budapest on July 9, 1944, in an effort to save Hungary's sole remaining Jewish community. Over the next six months, he worked with other neutral diplomats, the Vatican, and Jews themselves to prevent the deportation of these last Jews. Elsewhere, Le Chambon-sur-Lignon, a French Huguenot village, became a haven for 5,000 Jews. In Poland, where it was illegal to aid Jews and where such action was punishable by death, the Zegota (Council for Aid to Jews) rescued a similar number of Jewish men, women, and children. Financed by the Polish government in exile and involving a wide range of clandestine political organizations, the Zegota provided hiding places, financial support, and forged identity documents.

Rescue efforts

Some Germans, even some Nazis, dissented from the murder of the Jews and came to their aid. The most famous was Oskar Schindler, a Nazi businessman, who had set up operations using involuntary labour in German-occupied Poland in order to profit from the war. Eventually, he moved to protect his Jewish workers from deportation to extermination camps. In all occupied countries, there were individuals who came to the rescue of Jews, offering a place to hide, some food, or shelter for days, weeks, or even for the duration of the war. Most of the rescuers did not see their actions as heroic but felt bound to the Jews by a common sense of humanity. Israel later recognized rescuers with honorary citizenship and commemoration at Yad Vashem, Israel's memorial to the Holocaust.

Jewish resistance. It is often asked why Jews did not make greater attempts at resistance. Principally they had no access to arms and were surrounded by native anti-Semitic populations who collaborated with the Nazis or condoned the elimination of the Jews. In essence the Jews stood alone against a German war machine zealously determined to carry out the "final solution." Moreover, the Nazis went to great lengths to disguise their ultimate plans. Because of the German policy of collective reprisal, Jews in the ghettos often hesitated to resist. This changed when the Germans ordered the final liquidation of the ghettos, and residents recognized the imminence of their death.

Jews resisted in the forests, in the ghettos, and even in the

death camps. They fought alone and alongside resistance groups in France, Yugoslavia, and Russia. As a rule, full-scale uprisings occurred only at the end, when Jews realized the inevitability of impending death. On April 19, 1943, nine months after the massive deportations of Warsaw's Jews to Treblinka had begun, the Jewish resistance, led by 24-year-old Mordecai Anielewicz, mounted the Warsaw Ghetto Uprising. In Vilna partisan leader Abba Kovner, recognizing the full intent of Nazi policy toward the Jews, called for resistance in December 1941 and organized an armed force that fought the Germans in September 1943. In March of that year, a resistance group led by Willem Arondeus, a homosexual artist and author, bombed a population registry in Amsterdam to destroy the records of Jews and others sought by the Nazis. At Treblinka and Sobibor uprisings occurred just as the extermination camps were being dismantled and their remaining prisoners were soon to be killed. This was also true at Auschwitz, where the Sonderkommando ("Special Commando"), the prisoner unit that worked in the vicinity of the gas chambers, destroyed a crematorium just as the killing was coming to an end in 1944.

Warsaw Ghetto Uprising

Concealment and liberation. By the winter of 1944–45, with Allied armies closing in, desperate SS officials tried frantically to evacuate the camps and conceal what had taken place. They wanted no eyewitnesses remaining. Prisoners were moved westward, forced to march toward the heartland of Germany. There were over 50 different marches from Nazi concentration and extermination camps during this final winter of Nazi domination, some covering hundreds of miles. The prisoners were given little or no food and water, and almost no time to rest or take care of bodily needs. Those who paused or fell behind were shot. In January 1945, just hours before the Red Army arrived at Auschwitz, the Nazis marched some 60,000 prisoners to Wodzisław and put them on freight trains to the camps at Bergen-Belsen, Gross-Rosen, Buchenwald, Dachau, and Mauthausen. Nearly one in four died en route.

In April and May of 1945, American and British forces en route to military targets entered the concentration camps in the west and caught a glimpse of what had occurred. Even though tens of thousands of prisoners had perished, these camps were far from the most deadly. Still, even for the battle-weary soldiers who thought they had already seen the worst, the sights and smells and the emaciated survivors they encountered left an indelible impression. At Dachau they came upon 28 railway cars stuffed with dead bodies. Conditions were so horrendous at Bergen-Belsen that some 28,000 inmates died after they were freed, and the entire camp had to be burned to prevent the spread of typhus. Allied soldiers had to perform tasks for which they were ill-trained: to heal the sick, comfort the bereaved, and bury the dead. As for the victims, liberation was not a moment of exultation. Viktor Frankl, a survivor of Auschwitz, recalled, "Everything was unreal. Unlikely as in a dream. Only later—and for some it was very much later or never—was liberation actually liberating."

The Allies, who had early and accurate information on the murder of the Jews, made no special military efforts to rescue them or to bomb the camps or the railroad tracks leading to them. They felt that only after victory could something be done about the Jewish situation. Warnings were issued, condemnations were made, plans proceeded to try the guilty after the war, but no concrete action was undertaken specifically to halt the genocide. An internal memo to U.S. Secretary of the Treasury Henry Morgenthau, Jr., from his general counsel in January 1944 characterized U.S. State Department policy as "acquiescence to the murder of the European Jews." In response Morgenthau helped spur the creation of the War Refugee Board, which made a late and limited effort to rescue endangered Jews, mainly through diplomacy and subterfuge.

The War Refugee Board

THE AFTERMATH

Displacement of the survivors. Although the Germans killed victims from several groups, the Holocaust is primarily associated with the murder of the Jews. Only the Jews were targeted for total annihilation, and their elimination was central to Hitler's vision of the "New Germany." The intensity of the Nazi campaign against the Jews continued unabated to the very end of the war and at points even took priority over German military efforts.

When the war ended, Allied armies found between seven and nine million displaced persons living outside their own countries. More than six million people returned to their native lands, but more than one million refused repatriation. Some had collaborated with the Nazis and feared retaliation. Others feared persecution under the new communist regimes. For the Jews, the situation was different. They had no homes to return to. Their communities had been shattered, their homes destroyed or occupied by strangers, and their families decimated and dispersed. First came the often long and difficult physical recuperation from starvation and malnutrition, then the search for loved ones lost or missing, and finally the question of the future.

Many Jews lived in displaced-persons camps. At first they were forced to dwell among their killers because the Allies did not differentiate on the basis of religion, merely by nationality. Their presence on European soil and the absence of a country willing to receive them increased the pressure on Britain to resolve the issue of a Jewish homeland in British-administered Palestine. Both well-publicized and clandestine efforts were made to bring Jews to Palestine. In fact, it was not until after the establishment of the State of Israel in May 1948 and the liberalization of American immigration laws in 1948 and 1949 (allowing the admission of refugees from Europe) that the problem of finding homes for the survivors was solved.

Trial of the perpetrators. Upon liberating the camps, many Allied units were so shocked by what they saw that they meted out spontaneous punishment to some of the remaining SS personnel. Others were arrested and held for trial. The most famous of the postwar trials occurred in 1945–46 at Nürnberg, the former site of Nazi Party rallies. There, the International Military Tribunal tried 22 major Nazi officials for war crimes, crimes against the peace, and a new category of crimes: crimes against humanity. This new category encompassed "murder, extermination, enslavement, deportation, and other inhumane acts committed against any civilian population . . . persecution on political, racial, or religious grounds . . . whether or not in violation of the domestic laws of the country where perpetrated." After the first trials, 185 defendants were divided into 12 groups, including physicians responsible for medical experimentation (but not so-called euthanasia), judges who preserved the facade of legality for Nazi crimes, *Einsatzgruppe* leaders, commandants of concentration camps, German generals, and business leaders who profited from slave labour. The defendants made up, however, a minuscule fraction of those who had perpetrated the crimes. In the eyes of many, their trials were a desperate, inadequate, but necessary effort to restore a semblance of justice in the aftermath of so great a crime. The Nürnberg trials established the precedent, later enshrined by international convention, that crimes against humanity are punishable by an international tribunal.

The Nürnberg trials

Over the ensuing half-century, additional trials further documented the nature of the crimes and had a public as well as a judicial impact. The 1961 trial in Jerusalem of Adolf Eichmann, who supervised the deportations of Jews to the death camps, not only brought him to justice but made a new generation of Israelis keenly aware of the Holocaust. The Auschwitz trials held in Frankfurt am Main, West Germany, between 1963 and 1976 increased the German public's knowledge of the killing and its pervasiveness. The trials in France of Klaus Barbie (1987) and Maurice Papon (1996–98) and the revelations of François Mitterrand in 1994 concerning his indifference toward Vichy France's anti-Jewish policy called into question the notion of French resistance and forced the French to deal with the issue of collaboration. These trials also became precedents as world leaders considered responses to other crimes against humanity in places such as Bosnia and Rwanda.

Legacy of the Holocaust. The defeat of Nazi Germany left a bitter legacy for the German leadership and people.



The public burning of "un-Germanic" books by members of the SA and university students in Berlin in May 1933.



In Nazi Germany, Jews were required to wear a yellow Star of David on their clothing.



"No Names," acrylic on canvas by Alice Lok Cahana, c. 1992-98. In the artist's traveling collection.



(Left) The graffiti on the window of a Jewish-owned business in Vienna reads: "You Jewish pig, may your hands rot off!"

(Top) Hulton Getty/Stone. (middle left) United States Holocaust Memorial Museum. (right) Alice Lok Cahana. (bottom left) Marschatek—United States Holocaust Memorial Museum



German troops execute a small group of Poles.



Roll call of Roma (Gypsy) prisoners at the Dachau concentration camp in Germany.



The closing of the homosexual gathering place Eldorado in Berlin, 1933.



"Smoke," [illegible] on linen by Samuel Bak, 1997. In the Pucker Gallery collection, Boston.



Members of the SS burn the bodies of gassed prisoners in the open air at Auschwitz II (Birkenau) in German-occupied Poland.

(Top left) Dokumentationsarchiv des Oesterreichischen Widerstandes/United States Holocaust Memorial Museum. (top right) Lydia Chagoll/United States Holocaust Memorial Museum. (middle) Landesbildstelle Berlin/United States Holocaust Memorial Museum. (bottom left) courtesy of the Pucker Gallery. (bottom right) United States Holocaust Memorial Museum.





"Arbeit Macht Frei," acrylic ink and mixed media on paper by Alice Lok Cahana, c. 1992–98. In the artist's traveling collection. The entrance to the Auschwitz I camp in German-occupied Poland bore the motto "Arbeit Macht Frei" ("Work Makes One Free").



Jewish children are deported from Marysin, in German-occupied Poland, to the Chelmno death camp.



Jewish men awaiting death in a gas van at the Chelmno death camp in German-occupied Poland.

(Top) Alice Lok Cahana, (bottom left) Jacob Igra/United States Holocaust Memorial Museum, (bottom right) Instytut Pamięci Narodowej/Institute of National Memory/United States Holocaust Memorial Museum



"Elegy III," oil on canvas by Samuel Bak, 1997. In the collection of the Pucker Gallery, Boston.



A Jewish survivor shows U.S. generals Dwight D. Eisenhower, Omar Bradley, and George S. Patton a pyre where the SS attempted to cremate corpses before evacuating the Buchenwald concentration camp in Germany, 1945.



Defendant Adolf Eichmann listens as the court declares him guilty on all counts at his trial in Jerusalem in 1961.



A memorial ceremony is held near the commemorative sculpture at the Baby Yar site in Ukraine.

(Top left) Courtesy of the Pucker Gallery, Boston. (top right) Harold Royall/United States Holocaust Memorial Museum. (bottom left) Babi Yar Society/United States Holocaust Memorial Museum. (bottom right) United States Holocaust Memorial Museum

Germans had committed crimes in the name of the German people. German culture and the German leadership—political, intellectual, social, and religious—had participated or been complicit in the Nazi crimes or been ineffective in opposing them. In an effort to rehabilitate the good name of the German people, the Federal Republic of Germany (West Germany) firmly established a democracy that protected the human rights of all its citizens and made financial reparations to the Jewish people in an agreement passed by parliament in 1953. West German democratic leaders made special efforts to achieve friendly relations with Israel. In the German Democratic Republic (East Germany), the communist leaders attempted to absolve their population of responsibility for the crimes, portraying themselves as the victims of the Nazis, and Nazism as a manifestation of capitalism. The first gesture of the post-communist parliament of East Germany, however, was an apology to the Jewish people. At one of its first meetings in the newly renovated Reichstag building in 1999, the German parliament voted to erect a Holocaust memorial in Berlin. The first state visitor to Berlin after its reestablishment as capital of a united Germany was Israeli Prime Minister Ehud Barak.

At the beginning of the 21st century, the history of the Holocaust continued to be unsettling. The Swiss government and its bankers had to confront their role as bankers to the Nazis and in recycling gold and valuables taken from the victims. Under the leadership of German prime minister Gerhard Schröder, German corporations and the German government established a fund to compensate Jews and non-Jews who worked in German slave labour and forced labour programs during the war. Insurance companies were negotiating over claims from descendants of policyholders killed during the war—claims that the companies denied immediately after the war by imposing prohibitive conditions, such as the presentation of a death certificate specifying the time and place of death of the insured. In several eastern European countries, negotiations addressed Jewish property that the Nazis had confiscated during the war but that could not be returned under the region's communist governments. Artworks stolen during the war and later sold on the basis of dubious records were the subject of legal struggles to secure their return to the original owners or to their heirs. The German government continued to pay reparations—first awarded in 1953—to individual Jews and the Jewish people to acknowledge responsibility for the crimes committed in the name of the German people.

Compensation

Artistic responses to the Holocaust. Artists the world over and camp survivors themselves have responded to the Holocaust through art. The very existence of Holocaust art can, however, create a sense of unease. Critic Irving Howe has asked, "Can imaginative literature represent in any profound or illuminating way the meanings of the Holocaust? Is 'the debris of our misery' (as one survivor described it) a proper or manageable subject for stories and novels? Are there not perhaps extreme situations beyond the reach of art?" Similarly, philosopher Theodore Adorno has commented that writing poetry after Auschwitz is barbaric. Yet poetry has been written—moving poetry that seeks to come to terms with the tragedy even in the German language—in works by Nelly Sachs and Paul Celan, among others. Gripping work dealing with the horror, pain, and loss of the Holocaust has appeared in every literary genre and in music, film, painting, and sculpture.

Literature of the Holocaust

Survivors of the Holocaust have produced powerful works that record or reflect on their experiences. Anne Frank's *The Diary of a Young Girl* (originally in Dutch, 1947), Eli Wiesel's *Night* (originally in Yiddish, 1956), and works by Primo Levi are some of the most memorable in the field of literature. Paintings and drawings by survivors Samuel Bak, Alice Lok Cahana, and David Olère document the horrors that they experienced in ghettos and death camps. Holocaust survivors have also composed a wide variety of music, including street songs, which gave voice to life in the ghetto; resistance songs, such as Hirsh Glik's "Song of the Partisans" (composed and first performed 1943, published 1953); and classical compositions, such as *Quartet for the End of Time* (first performed 1941) by Olivier Messiaen and the opera *Der Kaiser von Atlantis oder der Tod dankt ab* (first performed 1943; "The Emperor of Atlantis or Death Abdicates") by Victor Ullman.

Artists of all kinds, regardless of any firsthand experience with the Holocaust, have sought to grapple with this tragedy. George Segal's memorial sculpture, *Holocaust*, is but one notable example. Visual art in response to the Holocaust includes paintings by Holocaust refugees Marc Chagall and George Grosz and the illustrated story *Maus* (published in installments 1980–85) by Art Spiegelman, the son of a survivor. Notable musical responses to the Holocaust include Arnold Schoenberg's *A Survivor from Warsaw* (first performed 1947), Dmitri Shostakovich's *13th Symphony* (first performed 1962), which used the text of the poem "Baby Yar" (1961) by Yevgeny Yevtushenko, and works by composers Charles Davidson, Michael Horvitz, and Oskar Morawetz.

Film, too, has been a prime medium for dealing with the Holocaust. Shortly after World War II, several eastern European filmmakers, including Aleksander Ford, Wanda Jakubowska, and Alfred Radok, attempted to capture the experience of Holocaust victims. Some of the most influential films since then include *The Diary of Anne Frank* (1959), directed by George Stevens; *Il Giardino dei Finzi-Contini* (1970, *The Garden of the Finzi Continis*), directed by Vittorio De Sica; the nine-hour documentary *Shoah* (1985), directed by Claude Lanzmann; *Au Revoir les Enfants* (1987, *Goodbye, Children*), directed by Louis Malle; *Schindler's List* (1993), directed by Steven Spielberg; *La Vita è Bella* (1997; *Life Is Beautiful*), directed by Roberto Benigni; and *Bent* (1997), directed by Sean Mathias and based on Martin Sherman's 1979 play about the Nazi persecution of homosexuals.

CONCLUSION

Today the Holocaust is viewed as the emblematic manifestation of absolute evil. Its revelation of the depths of human nature and the power of malevolent social and governmental structures has made it an essential topic of ethical discourse in fields as diverse as law, medicine, religion, government, and the military.

Many survivors report they heard a final plea from those who were killed: "Remember! Do not let the world forget." To this responsibility to those they left behind, survivors have added a plea of their own: "Never again." Never for the Jewish people. Never for any people. They hope that remembrance of the Holocaust can prevent its recurrence. In part because of their efforts, interest in the event has increased rather than diminished with the passage of time and in fact Holocaust Remembrance days are observed each year in many countries. More than half a century after the Holocaust, institutions, memorials, and museums continue to be built and films and educational curricula created to document and teach the history of the Holocaust to future generations.

BIBLIOGRAPHY

General references and histories: ISRAEL GUTMAN (ed.), *Encyclopedia of the Holocaust*, 4 vol. (1990, reissued 4 vol. in 2, 1995), is a comprehensive and authoritative reference work. A useful reference on the geographic extent of the Holocaust is the UNITED STATES HOLOCAUST MEMORIAL MUSEUM, *Historical Atlas of the Holocaust* (1996). MICHAEL BERENBAUM (ed.), *Witness to the Holocaust* (1997), contains 94 basic documents on 21 major themes, from the Nazi rise to power to the Nürnberg trials. MICHAEL R. MARRUS, *The Holocaust In History* (1987, reissued 1989), offers insights on a variety of historical debates surrounding the Holocaust. MICHAEL BERENBAUM, *The World Must Know: The History of the Holocaust as Told in the United States Holocaust Memorial Museum* (1993) is a non-technical, illustrated history of the Holocaust. Other general histories of the Holocaust include MARTIN GILBERT, *The Holocaust: A History of the Jews of Europe During the Second World War* (1986); RAUL HILBERG, *The Destruction of the European Jews*, rev. and definitive ed., 3 vol. (1985); LENI YAHIL, *The Holocaust: The Fate of European Jewry, 1932–45* (1990; originally published in Hebrew, 1987); and SAUL FRIEDLÄNDER, *Nazi Germany and the Jews*, vol. 1, *The Years of Persecution, 1933–39* (1997), the first of two planned volumes.

The perpetrators: For first-hand accounts of the Holocaust from the viewpoint of perpetrators and bystanders, see ERNST KLEE, WILLI DRESSEN, and VOLKER RIESS (eds.), *"The Good Old*

Days": The Holocaust as Seen by Its Perpetrators and Bystanders, trans. from German (1991; also published as *"Those Were the Days": The Holocaust through the Eyes of the Perpetrators and Bystanders,* 1993). HELEN FEIN, *Accounting for Genocide: National Responses and Jewish Victimization During the Holocaust* (1979, reprinted 1984), presents a sociological account of genocide and the social forces that make it possible. DANIEL JONAH GOLDHAGEN, *Hitler's Willing Executioners: Ordinary Germans and the Holocaust* (1996), is a controversial work exploring the rise of anti-Semitism in Germany and the complicity of ordinary Germans in the Holocaust. For an account of the human impact of the killing process in one *Einsatzgruppe,* see CHRISTOPHER R. BROWNING, *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland* (1992, reissued 1998). HENRY FRIEDLANDER, *The Origins of Nazi Genocide: From Euthanasia to the Final Solution* (1995), traces the development of genocidal policies and techniques in the Nazi T4 Program. GITTA SERENY, *Into that Darkness* (1974, reprinted 1991), offers a chilling account of prison interviews with Franz Stangl, commandant of Sobibor and Treblinka and a product of the German T4 camps. ROBERT JAY LIFTON, *The Nazi Doctors: Medical Killing and the Psychology of Genocide* (1986), explores the role and psychology of Nazi physicians. Biographies of Nazi architects of the Holocaust include RICHARD BREITMAN, *The Architect of Genocide: Himmler and the Final Solution* (1991), and ALAN BULLOCK, *Hitler: A Study in Tyranny,* completely rev. ed. (1962, reissued 1995), also published in an abridged ed. with the same title (1971, reissued 1991). JOHN LUKACS, *The Hitler of History* (1997), is less a biography of Hitler and more a review of the way in which historians have treated him.

The victims: ISAIAH TRUNK, *Judenrat: The Jewish Council in Eastern Europe under Nazi Occupation* (1972, reissued 1996), describes the dilemma facing the Jewish Councils in the ghettos in their efforts to reconcile Jewish needs with Nazi demands. For an account of ghetto life and Jewish resistance to German aggression, see LUCY S. DAWIDOWICZ, *The War Against the Jews, 1933–1945,* 10th anniversary ed. (1986, reissued 1990). YISRAEL GUTMAN (Israel Gutman) and MICHAEL BERENBAUM (eds.), *Anatomy of the Auschwitz Death Camp* (1994, reissued 1998), a collection of essays, considers Auschwitz in context, each of its victim groups, and the inner life of both perpetrators and victims. TERRENCE DES PRES, *The Survivor: An Anatomy of Life in the Death Camps* (1976, reissued 1980), considers the experience of extermination camp inmates from a psychological viewpoint. LAWRENCE L. LANGER, *Holocaust Testimonies: The Ruins of Memory* (1991) explores the power of Holocaust survivors' testimonies and memories of their experience. Important firsthand accounts by Holocaust survivors include PRIMO LEVI, *If This Is a Man* (1959; originally published in Italian, 1947; also published as *Survival in Auschwitz,* 1961, reissued 1996), and ELIE WIESEL, *Night* (1960, reissued 1986; originally published in Yiddish, 1956).

Special topics: Edited volumes containing essays on different aspects of the Holocaust include JOHN K. ROTH and MICHAEL BERENBAUM (eds.), *The Holocaust: Religious and Philosophical Implications* (1989), on religious and philosophical issues related to the Holocaust; LAWRENCE L. LANGER (ed.), *Art from the Ashes* (1995), presenting art and literature on the Holocaust; CAROL RITTNER and JOHN K. ROTH (eds.), *Different Voices: Women and the Holocaust* (1993), on the issue of gender and women's experience of the Holocaust. Works on U.S. government policy on the Holocaust include HENRY L. FEINGOLD, *The Politics of Rescue: The Roosevelt Administration and the Holocaust, 1938–1945,* expanded and updated ed. (1980), a careful historical review; and DAVID S. WYMAN, *The Abandonment of the Jews* (1984, reissued 1998), a more critical indictment. TIM COLE, *Selling the Holocaust: From Auschwitz to Schindler, How History Is Bought, Packaged, and Sold* (1999); and NORMAN G. FINKELSTEIN, *The Holocaust Industry: Reflection on the Exploitation of Jewish Suffering* (2000), deal with what some people see as the commercialization of Holocaust remembrance.

(Mi.Bc.)

The Homeric Epics

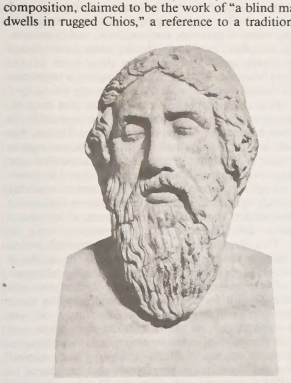
The two great epic poems of ancient Greece, the *Iliad* and the *Odyssey*, have always been attributed to a shadowy figure by the name of Homer. Little is known of him beyond the fact that his was the name attached in antiquity by the Greeks themselves to the two great poems. That there was an epic poet called Homer and that he played the primary part in shaping the *Iliad* and the *Odyssey*—so much may be said to be probable. If this assumption is accepted, then Homer must assuredly be one of the greatest of the world's literary artists.

Homer's influence

He is also one of the most influential authors in the widest sense, for the two epics provided the basis of Greek education and culture throughout the classical age and formed the backbone of humane education down to the time of the Roman Empire and the spread of Christianity. Indirectly through the medium of Virgil's *Aeneid* (which was loosely molded after the patterns of the *Iliad* and the *Odyssey*), directly through their revival under Byzantine culture from the late 8th century AD onward, and subsequently through their passage into Italy with the Greek scholars who fled westward from the Ottomans, the Homeric epics had a profound impact on the Renaissance culture of Italy. Since then the proliferation of translations has helped to make them the most important poems of the classical European tradition.

It was probably through their impact on classical Greek culture itself that the *Iliad* and the *Odyssey* most subtly affected Western standards and ideas. The Greeks regarded the great epics as something more than works of literature; they knew much of them by heart, and they valued them not only as a symbol of Hellenic unity and heroism but also as an ancient source of moral and even practical instruction.

Early references. Implicit references to Homer and quotations from the poems date to the middle of the 7th century BC. Archilochus, Alcman, Tyrtaeus, and Callinus in the 7th century and Sappho and others in the early 6th adapted Homeric phraseology and metre to their own purposes and rhythms. At the same time scenes from the epics became popular in works of art. The pseudo-Homeric "Hymn to Apollo of Delos," probably of late 7th-century composition, claimed to be the work of "a blind man who dwells in rugged Chios," a reference to a tradition about Homer himself. The idea that Homer had descendants known as "Homeridae," and that they had taken over the preservation and propagation of his poetry, goes back at least to the early 6th century BC. Indeed, it was not long before a kind of Homeric scholarship began: Theagenes of Rhegium in southern Italy toward the end of the same century wrote the first of many allegorizing interpretations. By the 5th century biographical fictions were well under way; the Pre-Socratic philosopher Heracleitus of Ephesus made use of a trivial legend of Homer's death—that it was caused by chagrin at not being able to solve some boys' riddle about catching lice—and the concept of a contest of quotations between Homer and Hesiod (after Homer the most ancient of Greek poets) may have been initiated in the Sophistic tradition. The historian Herodotus assigned the formulation of Greek theology to Homer and Hesiod and claimed that they could have lived no more than 400 years before his own time, the 5th century BC. This should be contrasted with the superficial assumption, popular in many circles throughout antiquity, that Homer must have lived not much later than the Trojan War about which he sang.



Restored bust thought to represent Homer, copied from a Greek original, *c.* 450 BC. In the Sala delle Muse, the Vatican.
By courtesy of the Vatican Museum

Personal history

The general belief that Homer was a native of Ionia (the central part of the western seaboard of Asia Minor) seems a reasonable conjecture for the poems themselves are in predominantly Ionic dialect. Although Smyrna and Chios early began competing for the honour (the poet Pindar, early in the 5th century BC, associated Homer with both), and others joined in, no authenticated local memory survived anywhere of someone who, oral poet or not, must have been remarkable in his time. The absence of hard facts puzzled but did not deter the Greeks; the fictions that had begun even before the 5th century BC were developed in the Alexandrian era in the 3rd and 2nd centuries BC (when false scholarship as well as true abounded) into fantastic pseudobiographies, and these were further refined by derivative scholars under the Roman Empire. The longest to have survived purports to be by Herodotus himself; but it is quite devoid of objective truth.

Modern inferences. Modern scholars agree with the ancient sources only about Homer's general place of activity. The most concrete piece of ancient evidence is that his descendants, the Homeridae, lived on the Ionic island of Chios. Yet an east Aegean environment is suggested for the main author of the *Iliad* by certain local references in the poem; that is, to the peak of Samothrace just appearing over the intervening mass of Imbros when seen from the plain of Troy, to the birds at the mouth of the Cayster near Ephesus, to storms off Icaria and northwest winds from Thrace. East Aegean colouring is fainter in the *Odyssey*, which is set primarily in western Greece; but the poem's vagueness over the position of Ithaca, for example, is not incompatible with the idea of a poet in Ionia elaborating materials derived from the farther side of the Greek world.

Admittedly, there is some doubt over whether the *Iliad* and the *Odyssey* were even composed by the same main author. Such doubts began in antiquity itself and depended mainly on the difference of genre (the *Iliad* being martial and heroic, the *Odyssey* picaresque and often fantastic), but they may be reinforced by subtle differences of vocabulary even apart from those imposed by different subjects. Aristotle's conception of the *Odyssey* as a work of Homer's old age is not impossible; but if the *Iliad* is the earlier of the two (as seems likely from its simpler structure and the greater frequency of relatively late linguistic forms in the *Odyssey*), then the *Odyssey* could have been created after its image, and as a conscious supplement, once the example of monumental composition had been given. In any case the similarities of the two poems are partly due

to the coherence of the heroic poetical tradition that lay behind both.

Date of the epics

The internal evidence of the poems is of some use in determining when Homer lived. Certain elements of the poetic language, which was an artificial amalgam never exactly reproduced in speech, indicate that the epics were not only post-Mycenaean in composition but also substantially later than the foundation of the first Ionian settlements in Asia Minor of about 1000 BC. The running together of adjacent short vowels and the disappearance of the semivowel digamma (a letter formerly existing in the Greek alphabet) are the most significant indications of this. At the other end of the time scale the development in the poems of a true definite article, for instance, represents an earlier phase than is exemplified in the poetry of the middle and late 7th century. Both stylistically and metrically, the Homeric poems appear to be earlier than the Hesiodic poems, which many scholars place not long after 700 BC. A different and perhaps more precise criterion is provided by datable objects and practices mentioned in the poems. Nothing, except for one or two probably Athenian additions, seems from this standpoint to be later than around 700; on the other hand, the role assigned in the *Odyssey* to the Phoenicians as traders, together with one or two other phenomena, suggests a date of composition—for the relevant contexts at least—of after 900. A few passages in the *Iliad* may imply a new form of fighting in close formation, dependent on the development of special armour for foot soldiers (hoplites) after about 750, and references to the Gorgon mask as a decorative motif point in the same direction. It is true that the poems contain many traditional and archaic elements, and their language and material background are a compound of different constituents originating at different dates. It seems, nonetheless, plausible to conclude that the period of composition of the large-scale epics (as distinct from their much shorter predecessors) was the 9th or 8th century, with several features pointing more clearly to the 8th. The *Odyssey* may belong near the end of this century, the *Iliad* closer to its middle. It may be no coincidence that cults of Homeric heroes tended to spring up toward the end of the 8th century, and that scenes from the epic begin to appear on pots at just about the same time.

Homer as an oral poet. But even if his name is known and his date and region can be inferred, Homer remains primarily a projection of the great poems themselves. Their qualities are significant of his taste and his view of the world, but they also reveal something more specific about his technique and the kind of poet he was. It has been one of the most important discoveries of Homeric scholarship, associated particularly with the name of an American scholar, Milman Parry, that the Homeric tradition was an oral one—that this was a kind of poetry made and passed down by word of mouth and without the intervention of writing. Indeed Homer's own term for a poet is *aoidos,* "singer." The *Odyssey* describes two such poets in some detail: Phemius, the court singer in the palace of Odysseus in Ithaca, and Demodocus, who lived in the town of the semi-mythical Phaeacians and sang both for the nobles in Alcinous' palace and for the assembled public at the games held for Odysseus. On this occasion he sings of the illicit love affair of Ares and Aphrodite in a version that lasts for exactly 100 Homeric verses. This and the other songs assigned to these singers—for example, that of the Trojan Horse, summarized in the *Odyssey*—suggest that ordinary *aoidoi* in the heroic tradition worked with relatively short poems that could be given completely on a single occasion. That is what one would expect, and it is confirmed by the habits of singers and audiences at other periods and in other parts of the world (the tradition of the poet-singers of Muslim Serbia has provided the most fruitful comparison so far). Whatever the favoured occasion for heroic song—whether the aristocratic feast, the religious festival, or popular gatherings in tavern or marketplace—a natural limitation on the length of a poem is imposed by the audience's available time and interest as well as by the singer's own physique and the scope of his repertoire. Such relatively short songs must have provided the backbone of the tradition inherited by Homer, and his portraits of Demodocus and Phemius are likely to be accurate in this respect. What Homer himself seems to have done is to introduce the concept of a quite different style of poetry, in the shape of a monumental poem that required more than a single hour or evening to sing and could achieve new and far more complex effects, in literary and psychological terms, than those attainable in the more anecdotal and episodic songs of his predecessors.

The poet as *aoidos*

Poetic techniques. It can be asked how one can be so confident in classing Homer himself as an oral singer, for if he differed from Phemius or Demodocus in terms of length, he may also have differed radically in his poetic techniques. The very nature of his verse may provide a substantial part of the answer. The style of the poems is "formulaic"; that is, they rely heavily on the use not only of stock epithets and repeated verses or groups of verses—which can also be found to a much lesser extent in a literate imitator like Virgil—but also on a multitude of fixed phrases that are employed time and time again to express a similar idea in a similar part of the verse. The clearest and simplest instance is the so-called noun-epithet formulas. These constitute a veritable system, in which every major god or hero possesses a variety of epithets from which the choice is made solely according to how much of the verse, and which part of it, the singer desires to use up. Odysseus is called divine Odysseus, many-counseled Odysseus, or much-enduring divine Odysseus simply in accordance with the amount of material to be fitted into the remainder of the hexameter (six-foot) verse. A ship is described as black, hollow, or symmetrical not to distinguish this particular ship from others but solely in relation to the qualities and demands of the rhythmical context. The whole noun-epithet system is both extensive and economical—it covers a great variety of subjects with very little exact reduplication or unnecessary overlap. It would seem that so refined and complex a system could not be the invention of a single poet but must have been gradually evolved in a long-standing tradition that needed both the extension and the economy for functional reasons—that depended on these fixed phrase units because of its oral nature, in which memory, practice, and a kind of improvising replace the deliberate, self-correcting, word-by-word progress of the pen-and-paper composer. Admittedly, the rest of Homer's vocabulary is not as markedly formulaic as its noun-epithet aspect (or, another popular example, as its expressions for beginning and ending a speech). Many expressions, many portions of sentences are individually invented for the occasion, or at least so it seems. Even so, there is a strongly formulaic and ready-made component in the artificial language that was used by Homer, including its less conspicuous aspects such as the arrangement of particles, conjunctions, and pronouns.

It looks, therefore, as though Homer must have trained as an ordinary *aoidos,* who began (like most of the present-day Yugoslav *guslari*) by building up a repertoire of normal-length songs acquired from already established singers. The greatest heroic adventures of the past must already have been prominent in any repertoire, especially the Panhellenic adventures of the Seven Against Thebes, the Argonauts, and the Achaean attack on Troy. Some aspects of the Trojan War might already have been expanded into songs of unusual length, though one that was still manageable on a single occasion. Yet the process was presumably carried much further in the making of the monumental *Iliad,* consisting of more than 16,000 verses, which would take four or five long evenings, and perhaps more, to perform. This breakthrough into the monumental, which made exceptional and almost unreasonable demands of audiences, presupposes a singer of quite exceptional capacity and reputation—one who could impose the new and admittedly difficult form upon his listeners by the sheer unfamiliar genius of his song. The 8th century BC was in other respects, too, an era of cultural innovation, not least in the direction of monumentality, and huge temples (like the early temple of Hera in Samos) and colossal funerary vases (like the mixing bowls and amphorae in the so-called Geometric style from the Dipylon cemetery in Athens) may have found a literary analogue in the idea of a vast poetical treatment of the Trojan War. But in an

The monumental poem

important sense Homer was building upon a tendency of all known oral heroic poetry toward elaboration and expansion. The singer does not acquire a song from another singer by simple memorization. He adjusts what he hears to his existing store of phrases, typical scenes, and themes, and he tends to replace what is unfamiliar to him with something he already knows, or to expand it by adding familiar material that it happens to lack. Every singer in a living oral tradition tends to develop what he acquires. There is an element of improvisation, as well as of memory, in his appropriation of fresh material; and judging by the practice of singers studied from the middle of the 19th century onward in Russia, Serbia, Cyprus, and Crete the inclination to adjust, elaborate, and improve comes naturally to all oral poets.

Typical elements of an oral poem

Cumulative poetic structure. Homer must have decided to elaborate his materials not only in quality but also in length and complexity. All oral poetry is cumulative in essence; the verse is built up by adding phrase upon phrase, the individual description by adding verse upon verse. The whole plot of a song consists of the progressive accumulation of minor motifs and major themes, from simple ideas (such as "the hero sets off on a journey" or "addresses his enemy") through typical scenes (such as assemblies of men or gods) to developed but standardized thematic complexes (such as episodes of recognition or reconciliation). Homer seems to have carried this cumulative tendency into new regions of poetry and narrative; in this as in other respects (for example, in his poetical language) he was applying his own individual vision to the fertile raw material of an extensive and well-known tradition.

The result is much more complex than with an ordinary traditional poem. Understanding the origin and essential qualities of the *Iliad* or the *Odyssey* entails trying to sort out not only the separate components of the pre-Homeric tradition but also Homer's own probable contributions, whether distinguishable by their dependence on the monumental idea or by their apparent novelty vis-à-vis the tradition as a whole or by other means. Dialectal and linguistic components must be identified as far as possible—survivals of the Mycenaean language, for example, or words used exclusively in the Aeolian cities of the west coast of Asia Minor, or Athenian dialect forms introduced into the poems after the time of Homer; so must specific references to armour, clothing, houses, burial customs, political geography, and so on, that are likely to be assignable to the Late Bronze Age, the Early Iron Age, or the period of Homer's own activity—at the very least to be taken as relatively early or late within the whole range of the poetic tradition down to Homer. These are the tasks of modern Homeric scholarship. Yet such different forms and ideas in Homer are not conveniently separated into distinct sections of the text, which can therefore be assigned to early or late phases of composition. On the contrary, they may coexist in a single (artificial) linguistic form or a single descriptive phrase. Any member of the tradition, not least Homer himself, may, moreover, have chosen to archaize on one occasion, to innovate on another. One result is that the epics are dubious authorities for the assessment of historical events like the attack on Troy or the status of workers, just as they are ambiguous sources for early Greek grammar or theology. Another is that they are not bound to a single worldview or period or mode of perception; rather, they unite judgments and experiences never seen together in "real" life into a whole that is literary but nevertheless revealing of the underlying structure of human existence.

Establishing Homer's contributions

Stabilizing the text. An important and difficult question, which affects the accuracy of modern Homeric texts, is that of the date when the epics became "fixed"—which means given authoritative written form, since oral transmission is always to some extent fluid. An alphabetic writing system reached Greece in the 9th or early 8th century BC; before that was a gap of 200 or 300 years, following the collapse of Mycenaean culture and the disappearance of Linear B writing (with each sign generally representing a syllable), during which Greece seems to have been nonliterate. During that interval, certainly, much of the epic tradition was formed. The earliest alphabetic inscriptions to have survived, a few of them containing brief scraps of hexameter verse, date from around 730 BC. Therefore, if Homer created the *Iliad* at some time after 750 BC, he could conceivably have used writing to help him. Some scholars think that he did. Others believe that he may have remained nonliterate (since literacy is not normally associated with oral creativity) but dictated the poem to a literate assistant. Still others believe that the poems may have been preserved orally and not too inaccurately at least until the middle years of the following, the 7th, century, when "literature" in the strict sense appeared in the poetry of Archiochus. There are objections to all three theories, but this much can be generally agreed: that the use of writing was in any case ancillary, that Homer behaved in important ways like a traditional oral poet. Some scholars are convinced that certain of the more subtle effects and cross-references of Homer's poetry would be impossible without the ability to consult a written text. That is doubtful; certainly the capacities even of ordinary oral poets in this direction are constantly surprising to habitual literates.

At least it may be accepted that partial texts of the epics were probably being used by the Homeridae and by professional reciters known as rhapsodes (who were no longer creative and had abandoned the use of the lyre) by the latter part of the 7th century BC. The first complete version may well have been that established as a standard for rhapsodic competitions at the great quadrennial festival at Athens, the Panathenaea, at some time during the 6th century BC. Even that did not permanently fix the text, and from then on the history of the epics was one of periodical distortion followed by progressively more effective acts of stabilization. The widespread dissemination of the poems consequent upon the growth of the Athenian book trade in the 5th century and the proliferation of libraries after the 4th was followed by the critical work of the Alexandrian scholar Aristarchus of Samothrace in the 2nd century BC, and much later by the propagation of accurate minuscule texts (notably the famous manuscript known as Venetus A of the *Iliad*), incorporating the best results of Greco-Roman scholarship, in the Byzantine world of the Middle Ages. Rare portions of either poem may have been added after, but not long after, the main act of composition; the night expedition that results in the capture of the Trojan spy Dolon and that fills the 10th book of the *Iliad,* some of the underworld scenes in the 11th book of the *Odyssey,* and much of the ending of the *Odyssey* after line 296 of the 23rd book (regarded by Aristarchus as its original conclusion) are the most probable candidates on the grounds of structure, language, and style.

Even apart from the possibilities of medium-scale elaboration, the *Iliad* and the *Odyssey* exemplify certain of the minor inconsistencies of all oral poetry, and occasionally the composer's amalgamation of traditional material into a large-scale structure shows through. Yet the overriding impression is one of powerful unity.

The Iliad. The *Iliad* is not merely a distillation of the whole protracted war against Troy but simultaneously an exploration of the heroic ideal in all its self-contradictoriness—its insane and grasping pride, its magnificent but animal strength, its ultimate if obtuse humanity. The poem is, in truth, the story of the wrath of Achilles, the greatest warrior on the Greek side, that is announced in its very first words; yet for thousands of verses on end Achilles is an unseen presence as he broods among his Myrmidons, waiting for Zeus's promise to be fulfilled—the promise that the Trojans will set fire to the Achaean ships and force King Agamemnon to beg him to return to the fight. Much of the poetry between the first book, in which the quarrel flares up, and the 16th, in which Achilles makes the crucial concession of allowing his friend Patroclus to fight on his behalf, consists of long scenes of battle, in which individual encounters alternate with mass movements of the opposing armies. The battle poetry is based on typical and frequently recurring elements and motifs, but it is also subtly varied by highly individualized episodes and set pieces: the catalog of troop contingents, the formal duels between Paris and Menelaus and Ajax and Hector, Helen's identifying of the Achaean princes, Agamemnon

The story of the *Iliad*

inspecting his troops, the triumph of Diomedes, Hector's famous meeting back in Troy with his wife Andromache, the building of the Achaean wall, the unsuccessful embassy to Achilles, the night expedition, Hera's seduction of Zeus and Poseidon's subsequent invigoration of the Achaeans. Patroclus' death two-thirds of the way through the poem brings Achilles back into the fight, although not before the recovery of Patroclus' body, the making of new divine armour for Achilles, and his formal reconciliation with Agamemnon. In book 22 he kills the deluded Hector; next he restores his heroic status by means of the funeral games for Patroclus; and in the concluding book Achilles is compelled by the gods to restore civilized values and his own magnanimity by surrendering Hector's body to King Priam.

The Odyssey. The *Odyssey* tends to be blander in expression and sometimes more diffuse in the progress of its action, but it presents an even more complex and harmonious structure than the *Iliad.* The main elements are the situation in Ithaca, where Penelope, Odysseus' wife, and their young son, Telemachus, are powerless before her arrogant suitors as they despair of Odysseus' return from the siege of Troy; Telemachus' secret journey to the Peloponnese for news of his father, and his encounters there with Nestor, Menelaus, and Helen; Odysseus' dangerous passage, opposed by the sea-god Poseidon himself, from Calypso's island to that of the Phaeacians, and his narrative there (from book 9 to book 12) of his fantastic adventures after leaving Troy, including his escape from the cave of the Cyclops, Polyphemus; his arrival back in Ithaca, solitary and by night, at the poem's halfway point, followed by his meeting with his protector-goddess Athena, his elaborate disguises, his self-revelation to the faithful swineherd Eumaeus and then to Telemachus, their complicated plan for disposing of the suitors, and its gory fulfillment. Finally comes the recognition by his faithful Penelope, his recounting to her of his adventures, his meeting with his aged father, Laertes, and the restitution, with Athena's help, of stability in his island kingdom of Ithaca. (See also GREEK LITERATURE: *Epic narrative.*)

The story of the *Odyssey*

Homer's influence seems to have been strongest in some of the most conspicuous formal components of the poems. The participation of the gods can both dignify human events and make them seem trivial—or tragic; it must for long have been part of the heroic tradition, but the frequency and the richness of the divine assemblies in the *Iliad,* or the peculiarly personal and ambivalent relationship between Odysseus and Athena in the *Odyssey,* probably reflect the taste and capacity of the main composer. The many-sidedness of battle, the equivocal realism of death in a hundred forms, must have been developed among Homer's predecessors but can never before have been deployed with such massive and complex effect. In the extended similes the strain of heroic action is relieved by the illuminating intrusion of a quite different and often peaceful contemporary world, in images developed often almost longingly beyond the immediate point of comparison. These similes, in their placing and their detail at least, surely depend on the main composer. And yet, beyond such general intuitions as these, the attempt to isolate his special contributions often becomes self-defeating. The *Iliad* and the *Odyssey* owe their unique status precisely to the creative and therefore unanalyzable confluence of tradition and design, the crystalline fixity of a formulaic style and the mobile spontaneity of a brilliant personal vision. "Homer" implies, above all, this fusion.

The result is an impressive amalgam of literary power and refinement. The *Iliad* and the *Odyssey,* however, owe their preeminence not so much to their antiquity and to their place in Greek culture as a whole but to their timeless success in expressing on a massive scale so much of the triumph and the frustration of human life. Although all literature must be engaged with that to some degree, epic poems are not where one most expects to find it. But these poems rise above the immediate concerns of heroic battle or the struggle against gods and nature or against monstrous forces, and they do so with the help of a poetical language of great simplicity and subtlety, a rugged and surprisingly variable narrative technique, and a nucleus of remarkable tales set around the Trojan War and its aftermath. Their greatest power lies, perhaps, in their dramatic quality because much of each poem consists of conversation and speeches, in which rhetoric is kept firmly under control and the individual characters emerge as they confront each other and the gods with advice, inquiry, request, resignation, and passion. Achilles, Hector, Menelaus, Ajax, Odysseus, and the others acquire a kind of heroic glow that even Greek tragedy later found hard to emulate. That is the result, in part, of the very archaism of these age-old tales, which the special techniques of monumental composition never attempted to conceal; but it also depends on something that overlaps that archaism, namely a sheer mythic quality imparting to these tales something of the universal validity to which all great literature aspires and which Homer achieved consistently and with an apparent ease that must be deceptive.

BIBLIOGRAPHY

Greek text: DAVID B. MONRO and THOMAS W. ALLEN (eds.), *Homeri Opera,* vol. 1–2, *Ilias,* 3rd ed. (1920, reprinted 1978), and vol. 3–4, *Odyssea,* 2nd ed. (1916, reprinted 1975–79).

Translations: The best close modern translations are RICHMOND LATTIMORE (trans.), *The Iliad of Homer* (1951, reprinted 1976), and *The Odyssey of Homer* (1967, reprinted 1975); and WALTER SHEWRING (trans.), *The Odyssey* (1980). A loose but powerful contemporary verse translation is ROBERT FITZGERALD (trans.), *The Iliad* (1974, reissued 1984).

Commentaries: G.S. KIRK (ed.), *The Iliad, a Commentary* (1985–), the first of a projected six-volume series; M.M. WILLCOCK (ed.), *The Iliad of Homer, Books I–XII* (1978), and *The Iliad of Homer, Books XIII–XXIV* (1984); and W.B. STANFORD (ed.), Ὁμήρου Ὀδύσσεια: *The Odyssey of Homer,* 2nd ed., 2 vol. (1958–59, reissued 1973–74). See also the Italian commentaries on the *Odyssey* by S. WEST (Books 1–4), J.B. HAINSWORTH (Books 5–8), A. HEUBECK (Books 9–12), and A. HOEKSTRA (Books 13–16) (1981–84).

Critical studies: W.A. CAMPS, *An Introduction to Homer* (1980); JASPER GRIFFIN, *Homer* (1980), an introduction for the general reader, and *Homer on Life and Death* (1980, reprinted 1983); HOWARD CLARKE, *Homer's Readers: A Historical Introduction to the "Iliad" and the "Odyssey"* (1981); NORMAN AUSTIN, *Archery at the Dark of the Moon: Poetic Problems in Homer's "Odyssey"* (1975); JAMES M. REDFIELD, *Nature and Culture in the "Iliad": The Tragedy of Hector* (1975); M.I. FINLEY, *The World of Odysseus,* 2nd rev. ed. (1977, reissued 1982); G.S. KIRK, *The Songs of Homer* (1962, reprinted 1977), abbreviated as *Homer and the Epic* (1965, reprinted 1974); G.S. KIRK (ed.), *The Language and Background of Homer: Some Recent Studies and Controversies* (1964, reissued 1967); ALBERT B. LORD, *The Singer of Tales* (1960, reissued 1978); PAUL MAZON, *Introduction à l'Iliade* (1943, reprinted 1967); DENYS L. PAGE, *History and the Homeric Iliad* (1959), and *The Homeric Odyssey* (1955, reprinted 1976); MILMAN PARRY, *L'Épithète traditionnelle dans Homère* (1928); ALAN J.B. WACE and FRANK H. STUBBINGS (eds.), *A Companion to Homer* (1962, reprinted 1974); T.B.L. WEBSTER, *From Mycenae to Homer* (1958, reprinted 1977); and W.J. WOODHOUSE, *The Composition of Homer's Odyssey* (1930, reprinted 1969).

Allied studies: JOHN CHADWICK, *The Decipherment of Linear B,* 2nd ed. (1968, reprinted 1970); M.I. FINLEY, *Early Greece: The Bronze and Archaic Ages,* rev. ed. (1981); and LORD WILLIAM TAYLOUR, *The Mycenaeans,* rev. ed. (1983).

(G.S.K.)

Hong Kong

Hong Kong (Wade-Giles romanization: Hsiang-kang; Pinyin: Xianggang) is a special administrative region of China located to the east of the Pearl River (Chu Chiang) estuary on the south coast of China. The region is bordered by Kwangtung province on the north and the South China Sea on the east, south, and west. It consists of Hong Kong Island, originally ceded by China to Great Britain in 1842, the southern part of the Kowloon Peninsula and Stonecutters (Ngong Shuen) Island (now joined to the mainland), ceded in 1860, and the New Territories, which include the mainland area lying largely to the north, together with 230 large and small offshore islands—all of which were leased from China for 99 years from 1898 to 1997. The Chinese-British joint declaration signed on Dec. 19, 1984, paved the way for the entire territory to be returned to China, which occurred July 1, 1997.

Hong Kong has 422 square miles (1,092 square kilometres) of land area, including land reclaimed from the sea, and the area continues to grow as more land is reclaimed. Hong Kong Island and its adjacent islets have an area of only about 35 square miles, while urban Kowloon, which includes the Kowloon Peninsula south of Boundary Street, and Stonecutters Island measure only about six square miles. The New Territories account for the rest of the area, amounting to more than 90 percent of the total. The Victoria urban district located on the barren rocks of the northwestern coast of Hong Kong Island is the place where the British first landed in 1841, and it has since been the centre of administrative and economic activities.

Hong Kong developed initially on the basis of its excellent natural harbour and the lucrative China trade, particularly opium dealing. It was the expansion of its territory, however, that provided labour and other resources necessary for sustained commercial growth that led to its becoming one of the world's major trade and financial centres. The community remains limited in space and natural resources, and it faces persistent problems of overcrowding, trade fluctuations, and social and political unrest. Nevertheless, Hong Kong has emerged strong and prosperous, albeit with a changed role, as an entrepôt, a manufacturing and financial centre, and a vital agent in the trade and modernization of China.

This article is divided into the following sections:

Physical and human geography 639
- The land 639
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 642
  - Ethnic composition
  - Linguistic composition
  - Religions
  - Demographic trends
- The economy 642
  - Resources
  - Agriculture and fishing
  - Industry
  - Finance
  - Trade
  - Transportation
- Administration and social conditions 643
  - Government
  - Education
  - Housing
  - Health and welfare
- Cultural life 644
  - Cultural milieu and the arts
  - Cultural institutions
  - Recreation
  - Press and broadcasting

History 644
- Early settlement 644
- Events before and during World War II 645
- Modern Hong Kong 645

Bibliography 645

Physical and human geography

THE LAND

Relief. In sharp contrast to the low-lying areas of the Pearl River delta, but conforming geologically and structurally to the great South China massif, a well-eroded upland region, Hong Kong has rugged relief and marked variations in topography. Structurally the area is an anticline, running northeast-southwest, that was formed toward the latter part of the Jurassic Period (about 150 million years ago). Lava poured into this structure and formed volcanic rocks that were later intruded by an extensive granitic dome. The harbour of Hong Kong was formed by the drowning of the denuded centre of the dome. The surrounding hills on the mainland and on Hong Kong Island are partly capped by volcanic rocks, and steep, scarplike concave slopes lead to the inner harbour.

Geologic characteristics

The area is a partially submerged, dissected upland terrain that rises abruptly to heights of more than 2,950 feet (900 metres); its backbone is made up of a series of ridges, running northeast to southwest, that tie in closely with the structural trend in South China. This trend is clearly observable from the alignment of Lantau Island and the Tolo Channel. From Mount Tai Mo—at 3,140 feet (957 metres) the highest peak in the territory—the series of ridges extends southwestward to Lantau Island, where the terrain rises to 3,064 feet on Lantau Peak and 2,851 feet on Sunset Peak. Extending southeastward from Mount Tai Mo, the Kowloon Peak attains an elevation of 1,975 feet, but there is an abrupt drop to about 650 feet at Devil's Peak. Victoria (Hong Kong) Harbour is well protected by mountains on Hong Kong Island that include Victoria Peak in the west, which rises to 1,810 feet, and Mt. Parker in the east, which reaches a height of about 1,742 feet.

Lowlands of the Hong Kong region, including floodplains, river valleys, and reclaimed land, occupy less than one-fifth of the land. Extensive lowland regions are found only north of Mount Tai Mo, in the Yuen Long and Sheung Shui plains. The urban area that spans the two sides of the harbour, with continued reclamation, takes up only about one-tenth of the level area.

Drainage. Hong Kong lacks a river system of any scope, the only exception being in the north where the Sham Chun River, which forms the boundary between Kwangtung and Hong Kong, flows into Deep Bay after collecting a number of small tributaries. Most of the streams are small, and they generally run perpendicular to the northeast-southwest trend of the terrain. The construction of reservoirs and their catchment systems has reduced the amount of fresh water available downstream.

Soils. In general, Hong Kong's soils are acidic and of low fertility. An exception is the alluvial soils, which are found mainly in the Deep Bay area, where the sediment-laden waters of the Pearl meet saline waters at high tide and slow down to deposit their sediments to form mud flats. Paddy rice farming and, more recently, inten-



MAP INDEX

Major divisions

Hong Kong Island . . . . . 22 15 N 114 11 E
Kowloon . . . . . 22 19 N 114 11 E
New Territories . . . 22 24 N 114 10 E

Urban areas

Aberdeen . . . . . 22 15 N 114 09 E
Fanling . . . . . 22 30 N 114 08 E
Hang Hau . . . . . 22 19 N 114 16 E
Kwai Chung . . . . . 22 22 N 114 06 E
Lo Wu . . . . . 22 32 N 114 07 E
Ma On Shan . . . . . 22 25 N 114 14 E
Rennie's Mill . . . . . 22 18 N 114 15 E
Sai Kung . . . . . 22 23 N 114 15 E
San Tin San Wai . . . 22 29 N 114 03 E
Sha Tin . . . . . 22 23 N 114 11 E
Sheung Shui . . . . . 22 31 N 114 07 E
Stanley . . . . . 22 13 N 114 12 E
Tai O . . . . . 22 15 N 113 51 E
Tai Po . . . . . 22 27 N 114 10 E
Tin Shui Wai . . . . . 22 28 N 114 00 E
Tseung Kwan O . . . 22 20 N 114 15 E
Tsuen Wan . . . . . 22 22 N 114 07 E
Tuen Mun . . . . . 22 24 N 113 58 E
Tung Chung . . . . . 22 17 N 113 56 E
Victoria . . . . . 22 17 N 114 09 E
Yuen Long . . . . . 22 26 N 114 02 E

Physical features and points of interest

Aberdeen Island . . . 22 14 N 114 09 E
Castle Peak Monastery . . . . 22 23 N 113 57 E
Chek Lap Kok Airport . . . . . 22 18 N 113 55 E
Chek Lap Kok Island . . . . . 22 18 N 113 56 E
Cheung Island . . . . 22 12 N 114 01 E
Chi Ma Bay Peninsula . . . . . 22 14 N 114 00 E
Chinese University of Hong Kong . . . 22 26 N 114 12 E
Chung Hom Bay, *archaeological site* . . . . . 22 13 N 114 12 E
Clear Water Bay . . 22 17 N 114 18 E
Crooked Harbour . . . . . 22 33 N 114 16 E
Crooked Island . . . 22 33 N 114 18 E
Deep Bay . . . . . 22 27 N 113 55 E
Discovery Bay . . . 22 18 N 114 01 E
Double Haven, *anchorage* . . . . 22 31 N 114 18 E
Double Island . . . . 22 31 N 114 18 E
East Lamma Channel . . . . . 22 14 N 114 09 E
Fan Lau Peninsula . . . . 22 12 N 113 51 E
Grass Island . . . . 22 29 N 114 22 E
Green Island . . . . 22 17 N 114 07 E
Ha Mei Bay . . . . . 22 12 N 114 07 E
Hei Ling Island . . . 22 15 N 114 02 E
High Island Reservoir . . . . . 22 23 N 114 21 E
High Island Reservoir . . . . . 22 23 N 113 59 E
Kadoorie Beach . . 22 23 N 113 59 E
Kai Keung Leng Peak . . . . . 22 28 N 114 05 E
Kap Shui Mun Bridge . . . . . 22 21 N 114 03 E
Kau Sai Island . . . 22 22 N 114 19 E
Kau Yi Island . . . . 22 17 N 114 04 E
Kowloon Peak . . . . 22 21 N 114 13 E
Kowloon Reservoirs . . . . . 22 21 N 114 09 E
Kwai Tau Leng Peak . . . . . 22 31 N 114 12 E
Lam Tong Channel . . . . . 22 15 N 114 15 E
Lamma Island . . . . 22 12 N 114 07 E
Lantau Channel . . . 22 10 N 113 50 E
Lantau Island . . . . 22 15 N 113 59 E
Lantau Peak . . . . . 22 15 N 113 55 E
Lantau South Country Park . . . . 22 13 N 113 51 E
Long Harbour . . . . 22 27 N 114 20 E
Ma On Peak . . . . . 22 25 N 114 15 E
Ma On Peak Country Park . . . . 22 24 N 114 15 E
Ma Wan Island . . . 22 21 N 114 03 E
Mai Po Marshes, *wildlife sanctuary* . . . . 22 30 N 114 03 E
Mirs Bay . . . . . 22 33 N 114 24 E
Nim Shue Bay, *archaeological site* . . . . . 22 18 N 114 01 E
Pak Tai To Yan Peak . . . . . 22 28 N 114 08 E
Pat Sin Leng Country Park . . . . 22 29 N 114 12 E
Pat Sin Range . . . . 22 29 N 114 13 E
Pearl Island . . . . . 22 22 N 113 59 E
Peng Island . . . . . 22 15 N 114 02 E
Plover Cove Country Park . . . . 22 30 N 114 16 E
Plover Cove Reservoir . . . . . 22 28 N 114 15 E
Po Lin Monastery . . . . 22 15 N 113 55 E
Po Toi Island Group . . . . . 22 11 N 114 16 E
Port Shelter, *bay* . . 22 21 N 114 17 E
Repulse Bay, *beach* . . . . . 22 14 N 114 12 E
Rocky Harbour . . . 22 20 N 114 19 E
Sai Kung East Country Park . . . . 22 24 N 114 21 E
Sai Kung West Country Park . . . 22 25 N 114 18 E
Sham Chun River . . 22 30 N 114 02 E
Sharp Peak . . . . . 22 26 N 114 22 E
Shek Pik Reservoir . . . . . 22 14 N 113 54 E
Shek Uk, Mount . . . 22 27 N 114 18 E
Shing Mun Reservoirs . . . . . 22 23 N 114 09 E
Silverstrand Beach . . . . . 22 20 N 114 16 E
Soko Islands . . . . 22 10 N 113 54 E
South China Sea . . . 22 13 N 114 20 E
Stonecutters Island, *peninsula* . . 22 19 N 114 08 E
Tai Lam Chung Reservoir . . . . . 22 23 N 114 01 E
Tai Lam Country Park . . . . . 22 24 N 114 03 E
Tai Long Bay . . . . 22 24 N 114 24 E
Tai Mo, Mount . . . 22 25 N 114 07 E
Tai Po Harbour . . . 22 26 N 114 12 E
Tai Tam Country Park . . . . . 22 14 N 114 13 E
Tai To Yan Peak . . 22 17 N 114 08 E
Tolo Channel . . . . 22 28 N 114 17 E
Trappist Monastery . . . . . 22 17 N 114 00 E
Tsing Ma Bridge . . 22 21 N 114 04 E
Tsing Yi Island . . . 22 21 N 114 05 E
Tung Lung Island . . 22 15 N 114 17 E
University of Hong Kong . . . . 22 17 N 114 08 E
Victoria Harbour . . 22 17 N 114 10 E
Victoria Peak . . . . 22 17 N 114 08 E
West Lamma Channel . . . . . 22 13 N 114 04 E
Yau Ma Tei Typhoon Shelter . . 22 19 N 114 10 E
Yim Tso Ha Egret Sanctuary . . . . . 22 32 N 114 12 E

sive vegetable cultivation have modified the alluvial soils. Elsewhere, hill soils, classified as red-yellow podzolic and krasnozem, abound. Under forest, these hill soils have a well-developed profile, with rich topsoil, but, when they are exposed, as is mostly the case, they tend to be thin and lacking in nutrients. Under tropical conditions, sheet and gully erosion is extensive and drastic.

Climate. Hong Kong lies at the northern fringe of the tropical zone. Its seasonal changes are well marked, however, with hot, humid summers and cool, dry winters. The climate is largely controlled by the pressure systems over the adjacent great Asian landmass and ocean surface. Thus, monsoonal winds blow from the northeast in winter as a result of the cooling of the landmass and the development of a large thermal anticyclone over Inner Mongolia. Southeast winds develop in summer when the North Pacific Ocean heats up more slowly through solar radiation and becomes a high-pressure area.

The mean January and July temperatures are about 60° F (16° C) and 84° F (29° C), respectively. The lowest recorded temperature was 32° F (0° C) in January 1893, and the highest was 97° F (36° C) in August 1900. Frost occasionally occurs on hilltops in winter. The average annual rainfall amounts to about 88 inches (2,220 millimetres), more than half of which falls during the summer months of June, July, and August; only about 10 percent falls from November to March. Tropical cyclones, or typhoons, generally occur between June and October, and, of the 20 to 30 typhoons formed over the western North Pacific and South China Sea each year, an average of five or six may affect Hong Kong. The torrential downpours and strong winds that frequently accompany the typhoons sometimes devastate life and property in Hong Kong and adjacent areas of Kwangtung.

Typhoons

Plant and animal life. Hong Kong is noted for the lushness and great diversity of its plant life. The transitional climate between humid subtropical and warm temperate maritime excludes the most sensitive humid tropical genera due to the cool, dry winter conditions, but many tropical as well as temperate-zone families are represented. Most of the land, except for the heavily eroded badlands, is under tropical herbaceous growth, including mangrove and other swamp cover. The most common forest genus today is *Pinus,* represented by the native South China red pine and the slash pine, introduced from Australia. Some of the oldest areas of woodland are in the *feng-shui* wood, or "sacred groves," found in many New Territories villages. These woods consist essentially of native forest trees, some of which are of potential value to the villagers. Centuries of cutting and burning have, however, destroyed much of Hong Kong's original vegetation, leaving only about one-eighth of the land forested. A large portion of Hong Kong's present-day forest cover owes its origin to postwar afforestation programs, which have restored some of the stands of pine, eucalyptus, banyan, casuarina, and palm trees.

Vegetation loss and afforestation

Hong Kong's animal life consists of a mixture of mammals adapted to the subtropical environment. Among the few arboreal mammals are two species of nonnative monkeys that flourish in forests of the New Territories, the rhesus macaque and the long-tailed macaque. Tigers are reputed to have once roamed the area, but they are no longer in evidence. The largest remaining carnivores are rare and include the South China red fox, the Chinese leopard cat, the seven-banded civet, and the masked palm civet. Some rat and mouse species typically inhabit scrubland and grassland areas. Birdlife is abundant, and there are numerous species of snakes, lizards, and frogs.

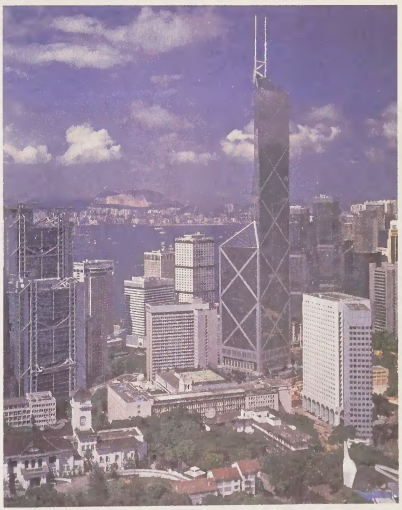
Settlement patterns. The predominantly urban settlements of Hong Kong are typically distributed linearly, following the irregular coastline and transportation routes. The principal urban areas are established on Hong Kong Island and Kowloon Peninsula, where more than half of the total population lives. There, most of the population is concentrated around Victoria Harbour, living on the limited flatland that is being continuously extended by reclamation. Many major streets, especially those on the northern shore of Hong Kong Island, as well as the entire industrial district of Kwun Tong and the eastern portion of Tsim Sha Tsui at the tip of the peninsula, have been built on reclaimed land.

In the New Territories north of the Kowloon hills, rural settlements vary from hamlets to small towns. Most of the villages, compactly built and often walled, follow the alignment of the river systems in the low-lying but fertile alluvial floodplains or the major route corridors. Villages of the Cantonese people are located mainly in the flat alluvial regions, whereas villages of the Hakka people usually are found in narrow valleys or on foothills. The *feng-shui* grove and pond are characteristic of both the Cantonese and Hakka villages; the grove is generally planted on the upslope, or back side, of a village for shade and protection, and the pond is for fish-farming.

A number of new towns have sprung up in the New Territories as a result of the tremendous increase in population there. Among these are Tsuen Wan, Tuen Mun (Castle Peak before 1973), and Sha Tin, which were established in the 1960s and designed to have populations of between 500,000 and 700,000 each. Others, including Tai Po, Fanling, Yuen Long, and, more recently, Tseung Kwon O (Junk Bay), were designated as new towns in the 1970s, with population goals ranging from about 150,000 to 400,000. Thus, the New Territories, where only one-eighth of the population resided in 1961, accounted for nearly half of the total by the mid-1990s, and some four-fifths of the New Territories population was concentrated in the new towns.

The new towns

True to its original character as a fishing port, Hong Kong has a sizable, though rapidly dwindling, marine settlement. The "boat people," or Tanka as they are locally known, are essentially fisherfolk living on junks and boats, as their ancestors did for centuries before them. They inhabit fishing towns such as Aberdeen, Shau Kei Wan, and Cheung Chau and typhoon shelters in the harbour areas. With the advance of urbanization and the decline of fishing activity, increasing numbers of them are working ashore.



The Bank of China Tower (centre right) in the Central District of Victoria, Hong Kong. Victoria Harbour and the Kowloon Peninsula are in the background.

© Paul Warchol

THE PEOPLE

Ethnic composition. The great majority of the population is Chinese by place of origin, the non-Chinese making up only about 2 percent of the total. Non-Chinese groups are split fairly evenly between non-Asians and Asians. British, Americans, Australians, Canadians, and New Zealanders are among the non-Asians, while the Asian minority groups include Japanese, Indians, Pakistanis, and Singaporeans. An overwhelming majority of the Chinese are from Kwangtung province and from Hong Kong itself, while less than 10 percent come from other parts of China, notably Fukien, Shanghai, Chekiang, and Kiangsu, and from Taiwan.

Linguistic composition. Chinese and English are both official languages. Chinese, especially Cantonese in the spoken form, is the common language, however, and is almost universally understood. A variety of dialects and other languages are used among the ethnic minorities. Apart from Cantonese, common dialects, such as Siyi, Chaochow, Hakka, Hoklo, and Tanka, are popularly used within separate native communities of the Kwangtung and Hong Kong Chinese. Groups from other parts of China are also likely to use their own native dialects, and, similarly, the non-Chinese are likely to use their own native languages among themselves. The use of Mandarin Chinese is rising as Hong Kong reintegrates with China.

Religions. The religious persuasions of the people of Hong Kong are as various as their languages and dialects. Among the Chinese, followers of Buddhism and Taoism by far outnumber other groups, and the numerous Buddhist and Taoist temples and monasteries, some centuries old, play an important role in the daily life of the average Chinese. Although each temple is generally dedicated to one or two deities, it is not unusual to find images of a number of other gods or goddesses inside. For a fishing and trading port, the most significant deities are those associated with the ocean and the weather, such as T'ien Hau, the goddess of heaven and protector of seafarers, who is honoured by temples at virtually every fishing harbour. Other leading deities include Kuan-yin (Avalokiteśvara), the Buddhist goddess of mercy; Hung Shing, god of the South Seas and a weather prophet; and Wong Tai Sin, a Taoist saint and deity. Christians constitute some half-million people, divided roughly equally between Roman Catholics and some 50 Protestant denominations and sects such as Anglican, Baptist, Lutheran, and Methodist. There are also small numbers of Muslims, Hindus, Sikhs, and Jews.

Significant deities



Vegetable farm on Pok Toi Chau (Lamma Island).
Bob Davis—Aspect Picture Library, London

Demographic trends. As in many large urban centres of the world, Hong Kong's population has increased in the late 20th century. Since the 1950s the average annual rate of growth has fluctuated between about 2 and 4 percent, the variations in some degree based on the sporadic flow of immigrants from China. Immigration has been a chief cause of population increase, but it was slowed through changes in immigration policy in 1980 and 1982, and emigration rose from the late 1980s. Birth rates have steadily declined since the late 1950s, the rate of natural increase falling below 1 percent by the 1980s. Life expectancy, however, has been showing a gradual increase. Since the 1950s, the proportion of the population under 15 has decreased rapidly, while that between 15 and 64 has shown a marked increase and the group over 65 has more than doubled. Hong Kong is one of the world's most densely populated places.

THE ECONOMY

With its limited natural resources, Hong Kong depends on imports for virtually all of its requirements, including raw materials, food and other consumer goods, capital goods, and fuel. Under its unique status as an international free port, entrepôt trade, mainly with China, flourished until 1951, when a United Nations embargo on trade with China and North Korea drastically curtailed it. This situation, combined with the need to export and with the availability of cheap labour, led to the establishment of competitive light industries and a transformation of the economy in the early 1960s. The market economy and the laissez-faire policy of the British colonial government provided flexibility for further industrialization and the incentive and freedom, from the late 1960s, to attract foreign investment and financial transactions. In succeeding years, with China adopting a more open foreign policy, entrepôt trade rapidly revived, while Hong Kong–China trade surged. Hong Kong developed not only in manufacturing, trade, and shipping but also as a regional financial centre and as an agent in China's pursuit of modernization. The tertiary (services) sector of the economy now makes up some four-fifths of the gross domestic product (GDP).

Stimulus of trade embargo

Resources. Hong Kong is practically devoid of any significant mineral resources. The mining for graphite and lead at Cham ("Needle") Hill and iron ore at Mount Ma On stopped long ago. Hong Kong is similarly poorly endowed in other natural resources: no commercial timber is produced from its sparse forest cover, and there is no hydroelectric potential from the small and short streams. Indeed, even water has been in serious short supply as a consequence of the limited areal extent, the steep terrain, and the lack of catchment areas. In spite of the many reservoirs, which were built mostly before World War II, and several giant projects, such as the water desalinization plant at Castle Peak and the Plover Cove and the High Island reservoirs, which are enclosed sea areas, the bulk of water consumed is piped in from Kwangtung province.

Agriculture and fishing. Only 6 percent of Hong Kong's land area is arable, and another 2 percent is under fishponds. Since the 1950s about one-third of the agricultural land has been lost to other uses. The growing season is year-round, however, and several crops a year are possible. Paddy rice cultivation once dominated agricultural land use, but it has practically disappeared, having been surpassed by vegetable and pond fish farming. Other minor uses include the production of fruits, flowers, and crops such as sweet potatoes, taro, yams, and sugarcane.

Marine fishing in the adjacent waters is one of Hong Kong's most important primary activities. Apart from pond fish, a marine fish culture has shown signs of development, notably in the eastern New Territories.

Industry. The rapid development of manufacturing in the 1950s was made possible by immigrant Chinese industrialists, mainly from Shanghai, who brought with them technology and capital. Foreign investments soon began to flow in to tap the huge supply of cheap labour and relatively cheap raw materials available in the surrounding

region. Most of the industry has been confined to the urban areas, especially in the densely populated districts of Kowloon. With the development of industrial and other new towns, manufacturing began to disperse into Kwun Tong, Tsuen Wan, Tuen Mun, and others. In 1977 the Hong Kong Industrial Estates Corporation was established to develop and manage industrial estates that would accommodate high-technology industries, first on reclaimed land in Tai Po and later in Yuen Long.

Manufacturing, once the most important sector of the Hong Kong economy, has been overshadowed by the vast service sector; it now constitutes less than one-tenth of the gross domestic product and employs about one-tenth of the labour force. Textile and clothing production is the leading manufacturing activity and contributes about one-third of the value of domestic exports. The electronics industry is the second largest export earner. There are some heavy industries such as shipbuilding and repair and aircraft engineering. Steel rolling, production of machine parts and plastics, and cement manufacturing serve local needs. The tourist industry, which is highly promoted by the government and well catered to by the huge service sector, is another significant part of the economy.

Manufactured products

Finance. Since 1969 Hong Kong has emerged as one of the major financial centres of the Asia-Pacific region, despite the fact that it is without the services of a central bank. The regional government delegates the functions of such an institution to certain government offices and selected commercial banks. In addition to the licensed banks in the region, there are representative offices of foreign banks, including registered deposit-taking companies.

Domestic and international currencies are traded at the Hong Kong foreign-exchange market. The stock market attracts investment from both foreign and domestic sources. Some of its major shares are also traded on the London stock market. A gold bullion market, once the world's largest, is operated by the Chinese Gold and Silver Exchange Society. The lack of exchange controls has contributed to the success of Hong Kong as a financial centre.

Trade. Hong Kong's free-trade policy has made the territory one of the world's great centres of trade. There is no tariff on imports, except for some luxury items, such as perfumes, motor vehicles, alcoholic beverages, and tobacco. Hong Kong is dependent upon imported products, which make up about half of the total amount of external trade, the rest being divided between exports and reexports. Apart from trade with other regions of China, Japan supplies the largest percentage of imports, and other major suppliers include Taiwan, the United States, Singapore, South Korea, Germany, and the United Kingdom. Machinery and transport equipment are the largest import categories, followed by consumer goods such as clothing, radios, television sets, stereos, computers, and watches. The third largest group includes raw materials and semi-manufactured goods such as synthetic and natural textiles, chemicals, and electronic components. Other significant imported products are foodstuffs and fuels.

China became the main market for Hong Kong's products prior to 1997, and this trade remained predominant after the territory's reintegration. Other major markets include the United States, the United Kingdom, Germany, Taiwan, and Japan. Textiles and clothing are the leading exports. Also important are electrical machinery and appliances, office machinery, photographic apparatus, and a variety of other manufactured items. Reexports constitute a major portion of the goods shipped out of Hong Kong, which developed historically as an entrepôt port.

Transportation. With roadways limited relative to the population, the government has enforced strict limitations on automobile ownership and placed heavy emphasis on the development of public transportation. As a result, the rate of car ownership is low, although it is steadily rising. The majority of the populace makes its daily trips by public transport. Apart from the bus, tram (streetcar), and ferry, the public is also served by a unique minibus service, a rapid transit system, and an electric railway. Buses, however, are the largest carrier, responsible for more than half of the daily public transport trips excluding those by taxi, followed by the combined minibus and maxicab (a regulated form of minibus) service. The precipitous Victoria Peak area is served by one of the oldest transport companies, which operates a cable car system between the peak and the Central District.

Public transport

International traffic is served by Hong Kong's international airport and its magnificent harbour, and there are good overland linkages with Kwangtung province. The new Hong Kong International Airport on Chek Lap Kok Island opened in 1998. The port of Hong Kong, based at one of the world's finest natural harbours, is renowned for its efficiency and capacity. The capacity of its container terminals at Kwai Chung ranks Hong Kong among the world's largest container ports. Speedy ferry service between Hong Kong and Macau and parts of Kwangtung is provided by various craft, including hydrofoils and hovercraft. Railroad transportation to Kwangtung is provided by the Kowloon-Canton Railway. Electrification of the railway and the growth, along its line, of the new towns of Sha Tin, Tai Po, and Fanling caused a considerable increase in passenger traffic. Externally, the line carries annually millions of tons of freight and head of livestock, as well as passenger traffic between Hong Kong and Kwangtung.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. When it was a colony, Hong Kong was administered by a governor, who was appointed by and represented the monarch of the United Kingdom, directed the government, served as the commander in chief, and presided over the two main organs of government, the Executive Council and the Legislative Council. With the resumption of Chinese sovereignty over the territory in July 1997, the Basic Law of the Hong Kong Special Administrative Region (promulgated by the National People's Congress of China in 1990) went into effect. The guiding principle of the Basic Law was the concept of "one country, two systems," under which Hong Kong was allowed to maintain its capitalist economy and to retain a large degree of political autonomy (except in matters of foreign policy and defense) for a period of 50 years.

The Basic Law vests executive authority in a chief executive, who is under the jurisdiction of the central government in Peking (Beijing) and serves a five-year term. Legislative authority rests with a Legislative Council, whose 60 members serve a four-year term; the chief executive, however, can dissolve the council before the end of a term. A 400-member Selection Committee, created by the central government, chose Tung Chee-hwa as the first chief executive. A Provisional Legislative Council was also chosen, which served for one year after reversion.

In May 1998, elections were held for the first formal Legislative Council (whose members served a two-year term), followed by elections for a full four-year term in September 2000. In each election, half the members were chosen by "functional constituencies," drawn from business and professional circles. In 1998, 20 members were elected from geographic constituencies, this number increasing to 24 in 2000. The remaining members—10 in 1998 and six in 2000—were chosen by an 800-member Election Committee, which was to cease functioning in subsequent elections. Ultimately, the chief executive and all council members are to be chosen by universal suffrage.

Civil and criminal law is derived generally from that of the United Kingdom, and the Basic Law states that this system is to be maintained. The highest court in the judiciary is the Court of Final Appeal, headed by a chief justice. This is followed by the High Court (headed by a chief judge) and by district, magistrate, and special courts. The chief executive appoints all judges, although judges of the Court of Final Appeal and the chief judge of the High Court also must be confirmed by the Legislative Council.

The judiciary

Education. Some three-fifths of the schools from kindergarten to secondary are private, and about another third are either subsidized or aided by public funds. The number of public schools in Hong Kong is quite small. Education is compulsory through the junior secondary level. Students finishing primary, junior secondary, and senior secondary education take examinations for allocation of school places at the next higher level.

Postsecondary schools, mostly subsidized, are strained by their small size, although efforts are being made to increase the size and number of institutions. The combined enrollment of the two universities, the University of Hong Kong (1911) and the Chinese University of Hong Kong (1963), nearly doubled between 1984 and 1994. The Hong Kong Polytechnic (1972), with mainly technical and vocational courses, and the Hong Kong Baptist University (1956), a private institution, offer degree courses in selected subjects. Colleges of education train teachers, mainly for primary schools, while responsibility for teacher training for secondary education rests with the two universities. In 1984 the City Polytechnic of Hong Kong (now the City University of Hong Kong) accepted its first students. Numerous other vocational, technical, and industrial training institutions operate throughout Hong Kong, and thousands of students are also enrolled in extension programs. Nonetheless, thousands also travel overseas each year for study. Nearly all of Hong Kong's people have at least a primary school education.

Housing. Historically, housing has been a major problem in Hong Kong, where space is limited and the number of occupants ever-growing. Changes in the residential environment between the establishment of the colony in 1842 and the Japanese occupation in 1941 were moderate, compared to those that took place in the postwar years. There was no planning in the earlier days of development, except that generally the British lived on the Peak (the area around Victoria Peak), other nationalities in the Mid-Levels (below the Peak), and the wealthy on somewhat higher ground, where the grand garden houses and large mansions remain as landmarks. Most of the Chinese lived on the lowlands surrounding the harbour, where the streets were narrow and the houses made of wood, bricks, and mortar. They lacked not only good natural lighting and ventilation but also piped water and flush toilets. Frequently urban development was the result of plagues, fires, and typhoons rather than comprehensive city planning; and the presence of large numbers of squatters and street sleepers as a result of a shortage of housing has been a persistent feature.

The limited housing supply was further reduced by the ravages of World War II. In the early postwar years, more than half of all families shared accommodations with others, living in cubicles, bed spaces, and attics and on roofs and verandas and in similar quarters. The colonial government's reluctant involvement in housing provision began with the building of resettlement blocks for fire victims in 1953, but it took real impetus in the early 1960s when the great demand for urban land resulted in the relocation of large numbers of squatters and urban poor. Eventually public housing came to accommodate more than half of the population, most of them living far from the urban core, and with growing numbers settling into the new towns. Land prices and rents have reached extremely high levels.

Public housing

Health and welfare. The health of the populace is generally good, the result, in part, of an aggressive program of public measures, including the promotion of preventive medicine and personal health services, and a relatively high quality of life. Improving health indexes and a downward trend in the occurrence of major communicable diseases are leading indicators of the state of health in Hong Kong. Most deaths are caused by cancer, heart disease, and cerebrovascular diseases. Hospitals are divided into three groups: government, government-assisted, and private. These are under great pressure to meet the needs of the people. Clinics, some operated by the government, supplement other medical facilities. Boat-borne clinics provide services to some outlying villages.

The social security system is largely limited to emergency relief programs, although there are some allowances for unemployment, old age, and disabilities. The aging of the population, coupled with the extreme crowding, presents increasing problems of elderly care.

CULTURAL LIFE

Cultural milieu and the arts. Hong Kong's is truly a mixed culture. Not only does the territory celebrate festivals and holidays of the East and the West, such as the Dragon Boat Festival, the Mid-Autumn Festival, the Lunar (Chinese) New Year, Christmas, the Western New Year, and others, but it also enjoys hundreds of annual cultural events ranging from traditional Cantonese and other Chinese regional operas and puppet shows to performances of ballet, theatre, and music and exhibitions of paintings and sculptures by nationally and internationally renowned performers and artists. The Hong Kong Arts Festival has become one of Asia's major cultural events, and the Hong Kong Philharmonic Orchestra, the Hong Kong Chinese Orchestra, the Chung Ying Theatre Company, and the City Contemporary Dance Company are among the best-known local artistic groups. The Hong Kong Conservatory of Music and the Hong Kong Academy of Ballet have been combined into the Hong Kong Academy for Performing Arts.

Music and theatre

Scores of motion pictures are produced every year in Hong Kong, many of which attain international fame; some have even started new trends in the art, such as the so-called kung fu films. The Hong Kong International Film Festival, inaugurated in 1977, is a major event, especially for the display of Asian films. Hong Kong is also a regional as well as an international centre in fashion design and in the cutting and design of ornamental diamonds.

Cultural institutions. Apart from the libraries of the major educational institutions, Hong Kong has a system of 25 libraries, including mobile ones. Of the museums, major ones include those specializing in history, art, science and technology, and space. The multifunctional City Hall (a cultural centre) and the Art Centre provide the major gallery, theatrical, and concert facilities. In addition, town halls have been established in the new towns and cultural centres in some districts to serve local communities.

Recreation. Hong Kong's country park system covers two-fifths of the land area, and outdoor recreation in parks is a part of the way of life for many of the people. City dwellers use park areas on the urban fringe for walking, running, and T'ai Chi ch'uan, among other activities, while remoter locations are used for kite flying, picnicking, hiking, cycling, and camping. There are well-organized programs of recreation and sports at the community level. The Ocean Park, one of the world's largest oceanariums, the Hong Kong Coliseum, a 12,500-seat indoor stadium that is among the largest in Asia, and the Queen Elizabeth Stadium are among the best venues for local and international sports events and musical, cultural, and entertainment programs. For those who can afford it, the many inlets and bays in Hong Kong provide a superb setting for pleasure sailing, waterskiing, canoeing, and other water sports.

Press and broadcasting. A wide-ranging and sophisticated communications network has developed in Hong Kong, reflecting its thriving commerce and international importance. There are some 60 newspapers (in various languages, but mostly Chinese) and the numbers of periodicals run into the hundreds. The territory is in addition the East and Southeast Asian headquarters for most of the major international news services. Broadcast news is provided by several television and radio companies, one of which is government-run. Under the British administration, the press developed largely free from government censorship. Television provides the major source of news and entertainment for the average family, and the Chinese television programs produced are not only for local consumption but also for overseas markets. Hong Kong also ranks as an important centre of publishing and printing: numerous books are published yearly for local consumption, several leading foreign publishers have their regional offices in Hong Kong, and many international magazines are printed in the territory.

For statistical data on Hong Kong, see the *Britannica World Data* section of the BRITANNICA BOOK OF THE YEAR.

History

EARLY SETTLEMENT

Archaeological remains of pottery, stone implements, rings, and bronzes found on more than 20 sites are

evidence of settlements in Neolithic times. In the 2nd millennium BC peoples are thought to have come to Hong Kong from North China. The Cantonese settled in the area about 100 BC; later came the Hakka, and by the mid-17th century the Hoklo had arrived. Hong Kong was the scene of the last struggles between the declining Ming dynasty and the rising Ch'ing, led by the Manchus.

Before the British arrived in the mid-19th century, Hong Kong Island had only a small fishing population, with little to recommend it for settlement. It lacked fertile soil and fresh water, was mountainous, and was reputed to be a haunt of pirates. But it was a relatively safe base for the British merchants who in 1821 began to use the fine harbour for the opium trade. The commercial and strategic value of this deep, sheltered harbour, possessing east and west entrances and lying on the main trade routes of Asia, was quickly realized.

After the first Opium War (1839–42), Hong Kong Island was ceded to Britain by the Treaty of Nanking. Not satisfied with only partial control of the harbour, the British forced China, under the Convention of Peking (1860) after the second Opium War (1856–60), to cede Kowloon Peninsula south of what is now Boundary Street and Stonecutters Island. By the Convention of 1898, the New Territories and 235 islands were leased to Britain for 99 years from July 1, 1898. With these additions, Hong Kong's population grew to 120,000 in 1861 and to more than 300,000 by the end of the century.

EVENTS BEFORE AND DURING WORLD WAR II

Since its establishment, Hong Kong, more than other treaty ports, afforded a refuge for people and capital from China as well as a way station for emigrants destined for Southeast Asia and beyond. Such movements between China and Hong Kong were free and were responsive to political and economic conditions in China. After the establishment of the Republic of China in 1912, early nationalists sought to abolish all foreign treaty privileges in China. A boycott of foreign goods particularly hurt Britain, which was well established in China. The campaign soon spread to Hong Kong, where strikes occurred in the 1920s.

When the Sino-Japanese War broke out in 1937, Hong Kong again became a refuge, with thousands of Chinese arriving before the advancing Japanese. With the outbreak of war in Europe in 1939, the position of the colony became even more precarious, and the Japanese attacked and occupied Hong Kong in December 1941. During the war Hong Kong's commerce was impaired; food was scarce, and many residents fled to inland China. The population, which numbered 1,600,000 in 1941, was about 650,000 in 1945 when the Japanese surrendered.

Japanese occupation

MODERN HONG KONG

British troops returned to the city on Aug. 30, 1945, and civil government was reestablished in May 1946. Hundreds of thousands of Chinese and foreigners returned, joined by refugees from China fleeing the civil war between the Nationalist and Communist armies.

The 1951 United Nations embargo on trade with China and North Korea during the Korean War curtailed the entrepôt trade, the colony's lifeline, and for several years conditions were depressed. Hong Kong began its revival with light industries such as textiles, which were set up by immigrant capitalists, provided needed employment, and became the basis for further industrialization. With much of the development relying on cheap labour, toiling under poor working conditions, labour disputes and social discontent began to spread in the early 1960s. Riots broke out in Hong Kong and Kowloon in May 1967 following a labour dispute in a plastic-flower factory. The economic and social unrest led to violent political demonstrations, largely inspired by followers of the Cultural Revolution (1966–76) in China. When the situation stabilized toward the end of the 1960s, working and living conditions were improved by labour legislation, government housing projects, and public works programs. Simultaneously, high-technology industries such as electronics were developed, and the property and financial markets prospered until early 1973, when the stock market collapsed as billions of dollars were drained out of Hong Kong. From the mid-1970s the economy resumed its upward trend as relations with China improved.

In the late 1970s, concern about the future of Hong Kong began to loom large, as British jurisdiction over the leased areas of the New Territories neared the 1997 expiration date. Although the lease applied only to the New Territories, the Chinese government had consistently maintained that the whole of Hong Kong was Chinese territory and that the earlier Hong Kong–British agreements were unequal treaties and were invalid. Initial contacts between the two governments on the matter began in March 1979, but negotiations did not start until late 1982 and continued for two years. Finally, the Chinese-British joint declaration concerning Hong Kong was signed by the heads of the two governments in Peking on Dec. 19, 1984. The agreement stipulated that Hong Kong (including Hong Kong Island, Kowloon, and the New Territories) would revert to China on July 1, 1997. Difficult negotiations ensued between Hong Kong and Peking on the final wording of the document by which Hong Kong would be governed under Chinese sovereignty. Despite some reservations from Hong Kong, the National People's Congress ratified the Basic Law on April 4, 1990, which took effect on July 1, 1997, and established the Hong Kong Special Administrative Region directly under the central government.

China-Britain agreement

The reversion has generally gone smoothly. The elimination of the colonial elected Legislative Council and the imposition of the provisional body was widely criticized, however. The central government's 1999 rejection of a judgment by Hong Kong's highest court on the rights of abode of children born on the mainland to Hong Kong residents also raised questions about the central government's respect for the principle of "one country, two systems." Hong Kong has gradually recovered from the Asian financial crisis of the late 1990s, continuing its shift from manufacturing to services and high tech industries.

For later developments in the history of Hong Kong, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics, in the *Macropædia* and *Micropædia*, see the *Propædia*, especially sections 96/10 and 975, and the *Index*.

BIBLIOGRAPHY

General works. A well-illustrated discussion of geography, history, economy, and society is presented in *Hong Kong* (annual), issued by the Hong Kong Government Information Service. ALAN BIRCH, Y.C. JAO, and ELIZABETH SINN (eds.), *Research Materials for Hong Kong Studies* (1984); and IAN SCOTT (compiler), *Hong Kong* (1990), are useful for further research on all aspects of Hong Kong.

Physical and human geography. Geologic studies of Hong Kong include P.M. ALLEN and E.A. STEPHENS, *Report on the Geological Survey of Hong Kong, 1967–1969* (1971). An analysis of the political, economic, geographic, and social developments in Hong Kong up to the early 1980s is found in CHI-KEUNG LEUNG, J.W. CUSHMAN, and WANG GUNGWU (eds.), *Hong Kong: Dilemmas of Growth* (1980). FRANK LEEMING, *Street Studies in Hong Kong: Localities in a Chinese City* (1977), examines Hong Kong's neighbourhoods. Socioeconomic studies include *Hong Kong Social and Economic Trends, 1970–1980* (1981), compiled by the Hong Kong Census and Statistics Department; A.J. YOUNGSON, *Hong Kong, Economic Growth and Policy* (1982); and SEK HONG NG and DAVID A. LEVIN, *Contemporary Issues in Hong Kong Labour Relations* (1983). *Hong Kong in Search of a Future* (1984), ed. by JOSEPH Y.S. CHENG; and NORMAN MINERS, *The Government and Politics of Hong Kong*, 5th ed., updated by JAMES T.H. TANG (1998), address the politics of the territory.

History. Overviews are provided by JAN MORRIS, *Hong Kong* (1988); NIGEL CAMERON, *An Illustrated History of Hong Kong* (1991); and MING K. CHAN and JOHN D. YOUNG (eds.), *Precarious Balance: Hong Kong Between China and Britain, 1842–1992* (1994). Aspects of the transition from British to Chinese rule are explored in ENBAO WANG, *Hong Kong, 1997: The Politics of Transition* (1995), an optimistic outlook; BRUCE BUENO DE MESQUITA, DAVID NEWMAN, and ALVIN RABUSHKA, *Red Flag Over Hong Kong* (1996), a more pessimistic forecast; and STEVE SHIPP, *Hong Kong, China: A Political History of the British Crown Colony's Transfer to Chinese Rule* (1995), which includes the complete texts of, among others, the 1984 Joint Declaration and the Basic Law.

(C.-K.L./Ed.)

Horses and Horsemanship

Zoologically, the horse is a mammal of the family Equidae. It comprises a single species, *Equus caballus*, whose numerous varieties are called breeds. Before the advent of mechanized vehicles, riding on horseback was one of the chief means of transportation. Horsemanship evolved, of necessity, as the art of riding with maximum discernment and a minimum of interference with the horse. Until the 20th century riding was a monopoly of the cavalry, of cowboys and others whose work required riding on horseback, and of the wealthy, who rode for sport. Although hunting and polo tend to remain sports of the wealthy and the role of the horse in battle has ended, special value is now placed on horse shows of a high standard, in which the most popular event is undoubtedly show jumping. Horsemanship has remained a valued social asset and symbol of prestige, but the opening of many new riding clubs and stables has made riding and horsemanship accessible to a much larger segment of the population. (C.E.C./Ed.)

This article is divided into the following sections:

General features of horses 646
- Form and function 646
  - Anatomical adaptations
  - Senses
  - Colour and pattern
  - Nutrition
  - Behaviour
  - Reproduction and development
  - Diseases and parasites
- Breeds of horses 648
  - Light horses
  - Heavy breeds
  - Ponies
- Evolution of the horse 650

Riding and horsemanship 651
- History 651
  - Origins and early history
  - Military horsemanship
- The art of horsemanship 651
  - Fundamentals
  - The horse's movements
  - Training and equipment
  - Seats
  - Dressage
  - Jumping
- Riding and shows 654
  - Horse shows
  - Olympic equestrian competition

Bibliography 655

General features of horses

Utility of the wild horse

In prehistoric times the wild horse was probably first hunted for food. When its domestication took place is unknown, but it certainly was long after the domestication of the dog or of cattle. It is supposed that the horse was first used by a tribe of Indo-European origin that lived in the steppes north of the chain of mountains adjacent to the Black and Caspian seas. Influenced by climate, food, and humans, the horse rapidly acquired its present form.

The relationship of the horse to humans has been unique. The horse is a partner and friend. It plowed fields and brought in the harvest, hauled goods and conveyed passengers, followed game and tracked cattle, and carried combatants into battle and adventurers to unknown lands. It has provided recreation in the form of jousts, tournaments, carousels, and the sport of riding. The influence of the horse is expressed in the English language in such terms as *chivalry* and *cavalier*, which connote honour, respect, good manners, and straightforwardness.

The horse is the "proudest conquest of Man," according to the French zoologist Le Comte de Buffon. Its place was at its master's side in the graves of the Scythian kings or in the tombs of the pharaohs. Many early human cultures were centred on possession of the horse. Superstition read meaning into the colours of the horse, and a horse's head suspended near a grave or sanctuary or on the gables of a house conferred supernatural powers on the place. Greek mythology created the centaur, the most obvious symbol of the oneness of horse and rider. White stallions were the supreme sacrifice to the gods, and the Greek general Xenophon recorded that "Gods and heroes are depicted on well-trained horses." A beautiful and well-trained horse was, therefore, a status symbol in ancient Greece. Kings, generals, and statesmen, of necessity, had to be horsemen. The names of famous horses are inseparably linked to those of their famous riders: Bucephalus, the charger of Alexander the Great; Incitatus, foolishly made a senator by the Roman emperor Caligula; El Morzillo, Cortés' favourite horse, to whom the Indians erected a statue; Roan Barbery, the stallion of Richard II, mentioned by Shakespeare; Copenhagen, the Duke of Wellington's horse, which was buried with military honours.

The horse in art

The horse has occupied a special place in the realm of art. From the Stone Age drawings to the marvel of the Parthenon frieze, from Chinese T'ang dynasty tomb sculptures to Leonardo da Vinci's sketches and Verrocchio's Colleoni, from the Qur'ān to modern literature, the horse has inspired artists of all ages and in all parts of the world.

The horse in life has served its master in travels, wars, and labours and in death has provided many commodities. Long before their domestication horses were hunted by primitive tribes for their flesh, and horsemeat is still consumed by people in parts of Europe and in Iceland and is the basis of many pet foods. Horse bones and cartilage are used to make glue. Tetanus antitoxin is obtained from the blood serum of horses previously inoculated with tetanus toxoid. From horsehide a number of articles are manufactured, including fine shoes and belts. The cordovan leather fabricated by the Moors in Córdoba, Spain, was originally made from horsehide. Stylish fur coats are made of the sleek coats of foals. Horsehair has wide use in upholstery, mattresses, and stiff lining for coats and suits; high-quality horsehair, usually white, is employed for violin bows. Horse manure, which today provides the basis for cultivation of mushrooms, was used by the Scythians for fuel. Mare's milk was drunk by the Scythians, the Mongols, and the Arabs.

FORM AND FUNCTION

A mature male horse is called a stallion, the female a mare. A stallion used for breeding is known as a stud. A castrated stallion is commonly called a gelding. Formerly, stallions were employed as riding horses, while mares were kept for breeding purposes only. Geldings were used for work and as ladies' riding horses. Recently, however, geldings generally have replaced stallions as riding horses. Young horses are known as foals; male foals are called colts and females fillies.

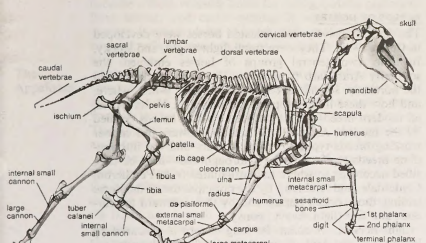
Anatomical adaptations. The primitive horse probably stood 12 hands tall (about 120 centimetres, or 48 inches) at the withers, the high point on the back at the base of the neck, and was dun coloured. Domestic horses gone wild, such as the mustangs of western North America, tend to revert to those primitive features under random mating; they generally are somewhat taller (about 15 hands), usu-

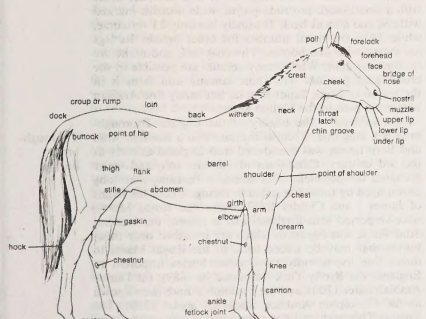
ally gray, dun, or brownish in colour, and move in herds led by a stallion.

The horse's general form is characteristic of an animal of speed: the long leg bones pivot on pulley-like joints that restrict movement to the fore and aft, the limbs are levered to muscle masses in such a way as to provide the most efficient use of energy, and the compact body is supported permanently on the tips of the toes, allowing fuller extension of the limbs in running (see illustration).

The rounded skull houses a large and complex brain, well developed in those areas that direct muscle coordination. While the horse is intelligent among subhuman animals, it is safe to say that the horse is more concerned with the functioning of its acute sensory reception and its musculature than with mental processes. Though much has been written about "educated" horses that appear to exhibit an ability to spell and count, it is generally agreed that in such cases a very perceptive animal is responding to cues from its master. But this ability is remarkable enough in its own right—for the cues are often given unconsciously by the human trainer, and detection of such subtle signals requires extremely sharp perception.

The horse, like other grazing herbivores, has typical adaptations for plant eating: a set of strong, high-crowned teeth, suited to grinding grasses and other harsh vegetation, and a relatively long digestive tract, most of which is intestine concerned with digesting cellulose matter from vegetation. Young horses have milk (or baby) teeth, which they begin to shed at about age two and a half. The permanent teeth, numbering 36 to 40, are completely developed by age four to five years. In the stallion these teeth are arranged as follows on the upper and lower jaws: 12 incisors that cut and pull at grasses; four canines, remnants without function in the modern horse and usually not found in mares; 12 premolars and 12 molars, high prisms that keep moving out of the jaw to replace the surfaces worn off in grinding food.





(Top) Skeleton of the horse; (bottom) points of the horse.

Under domestication the horse has diversified into three major types, based on size and build: draft horses, heavy limbed and up to 20 hands high (200 centimetres, or 80 inches); ponies, by convention horses under 14.2 hands high (about 144 centimetres, or 57 inches); and light horses—the saddle or riding horses—which fall in the intermediate size range. Domestic horses tend to be near-sighted, less hardy than their ancestors, and often high-strung, especially thoroughbreds, where intensive breeding has been focused upon speed to the exclusion of other qualities. The stomach is relatively small, and, since much vegetation must be ingested to maintain vital processes, foraging is almost constant under natural conditions. Domestic animals are fed several (at least three) times a day in quantities governed by the exertion of the horse.

Diversification into three types

Senses. The extremely large eyes placed far back on the elongated head admirably suit the horse for its chief mode of defense: flight. Its long neck and high-set eyes, which register a much wider range than do the eyes of a human being, enable the horse to discern a possible threat even while eating low grasses. Like human vision, the horse's vision is binocular, but only in the narrow area directly forward, and evidence suggests that it does not register colour. While visual acuity is high, the eyes do not have variable focus, and objects at different distances register only on different areas of the retina, which requires tilting movements of the head. The senses of smell and hearing seem to be keener than in human beings. As the biologist George Gaylord Simpson has put it in *Horses* (1961): "Legs for running and eyes for warning have enabled horses to survive through the ages, although subject to constant attack by flesh eaters that liked nothing better than horse for supper."

Colour and pattern. From the dun of the primitive horse has sprung a variety of colours and patterns, some highly variable and difficult to distinguish. Among the most important colours are black, bay, chestnut (and sorrel), palomino, cream, and white.

The black colour is a true black, although a white face marking (blaze) and white ankles (stockings) may occur. The brown horse is almost black but has lighter areas around the muzzle, eyes, and legs. Bay refers to several shades of brown, from red brown and tan to sandy. Bay horses have a black mane, tail, and (usually) stockings. Chestnut is similar to bay but with none of the bay's black overtones. Lighter shades of chestnut are called sorrel. The palomino horse runs from cream to bronze, with a flaxen or silvery mane and tail. The cream is a diluted sorrel, or very pale yellow, nearly white. White in horses is variable, ranging from aging grays (see below) to albinos with blue eyes and pink skin and to pseudoalbinos with a buff mane or with brown eyes. The chief patterns of the white horse are gray, roan, pinto, and appaloosa. Gray horses are born dark brown or black and develop white hairs as they age, becoming almost all white in advanced years. Roan refers to white mixed with other colours at birth: blue roan is white mixed with black; red roan is mixed white and bay; and strawberry roan is white and chestnut. The pinto is almost any spotted pattern of white and another colour; other names, such as paint, calico, piebald, skewbald, overo, and tobiano, refer to subtle distinctions in type of colour or pattern. Appaloosa is another extremely variable pattern, but the term generally refers to a large white patch over the hips and loin, with scattered irregular dark spots.

Nutrition. The horse's natural food is grass. For stabled horses, the diet generally consists of hay and grain. The animal should not be fed immediately before or after work, to avoid digestive problems. Fresh water is important, especially when the horse is shedding its winter coat, but the animal should never be watered when it is overheated after working. Oats provide the greatest nutritional value and are given especially to foals. Older horses, whose teeth are worn down, or those with digestive troubles, can be provided with crushed oats. Chaff (minced straw) can be added to the oat ration of animals that eat greedily or do not chew the grain properly. Crushed barley is sometimes

substituted in part for oats. Hay provides the bulk of the horse's ration and may be of varying composition according to locale. Mash is bran mixed with water and with various invigorating additions or medications. It may be given to horses with digestive troubles or deficient eating habits. Corn (maize) is used as a fattening cereal, but it makes the horse sweat easily. Salt is needed by the horse at all times and especially when shedding. Bread, carrots, and sugar are tidbits often used to reward an animal by the rider or trainer. In times of poverty horses have adapted to all sorts of food—potatoes, beans, green leaves, and in Iceland even fish—but such foods are not generally taken if other fare is available. A number of commercial feed mixes are available to modern breeders and owners; these mixes contain minerals, vitamins, and other nutrients and are designed to provide a balanced diet when supplemented with hay.

Behaviour. The horse's nervous system is highly developed and gives proof to varying degrees of the essential faculties that are the basis of intelligence: instinct, memory, and judgment. Foals, which stand on their feet a short while after birth and are able to follow their mothers within a few hours, even at this early stage in life exhibit the traits generally ascribed to horses. They have a tendency to flee danger. They express fear sometimes by showing panic and sometimes by immobility. Horses rarely attack and do so either when flight is impossible or when driven to assault a person who has treated them brutally.

Habit and instinct

Habit governs a large number of their reactions. **Instinct**, together with a fine sense of smell and hearing, enables them to sense water, fire, even distant danger. An extremely well-developed sense of direction permits the horse to find its way back to its stables even at night or after a prolonged absence. The visual memory of the horse prompts it to shy repeatedly from an object or place where it had earlier experienced fear. The animal's auditory memory, which enabled ancient army horses or hunters to follow the sounds of the bugles, is used in training. When teaching, the instructor always uses the same words and the same tone of voice for a given desired reaction. Intelligent horses soon attach certain movements desired by their trainers to particular sounds and even try to anticipate their rider's wishes.

While instinct is an unconscious reaction more or less present in all individuals of the same species, the degree of its expression varies according to the individual and its development. Most horses can sense a rider's uncertainty, nervousness, or fear and are thereby encouraged to disregard or even deliberately disobey the rider. Highbred animals, which give evidence of greater intelligence than those of low breeding, are capable not only of acts of vengeance and jealousy against their riders but also of expressions of confidence, obedience, affection, and fidelity. They are less willing than a lowbred horse to suffer rough handling or unjust treatment.

Cunning animals have been known to employ their intelligence and physical skill to a determined end, such as opening the latch of a stall or the lid of a chest of oats.

Reproduction and development. The onset of adult sex characteristics generally begins at the age of 16 to 18 months; the horse is considered mature, according to the breed, at approximately three years and adult at five. Fecundity varies according to the breed and may last beyond age 20 with thoroughbreds and to 12 or 15 with other horses. The gestation period is 11 months, 280 days being the minimum in which the foal can be born with expectation to live. As a rule a mare produces one foal per mating, twins occasionally, and triplets rarely. The foal is weaned at six months.

The useful life of a horse varies according to the amount of work it is required to do and the maintenance furnished by its owner. A horse that is trained carefully and slowly and is given the necessary time for development may be expected to serve to an older age than a horse that is rushed in its training. Racehorses that enter into races at the age of two rarely remain on the turf beyond eight. Well-kept riding horses, on the contrary, may be used more than 20 years.

The life span of a horse is calculated at six to seven times the time necessary for his physical and mental development; that is, 30 to 35 years at the utmost, the rule being about 20 to 25 years. Ponies generally live longer than larger horses. There are a number of examples of horses that have passed the usual limit of age. The veterinary university of Vienna conserves the skeleton of a thoroughbred mare of 44 years of age. There have been reports made of horses living to the early 60s in age.

Life expectancy

Diseases and parasites. Horses are subjected to a number of contagious diseases, such as influenza, strangles, glanders, equine encephalomyelitis, and swamp fever. Their skin is affected by parasites, including certain mites, ticks, and lice. Those with sensitive skin are especially subject to eczemas and abscesses, which may result from neglect or contamination. Sores caused by injuries to the skin from ill-fitting or unclean saddles and bridles are common ailments. The horse's digestive tract is particularly sensitive to spoiled feed, which causes acute or chronic indigestion, especially in hot weather. Worms can develop in the intestine and include the larvae of the botfly, pinworms, tapeworms, and roundworms (ascarids). Overwork and neglect may predispose the horse to pneumonia and rheumatism. The ailment known as roaring is an infection of the larynx that makes the horse inhale noisily; a milder form causes the horse to whistle. Chronic asthma, or "broken wind," is an ailment that is considered to be all but incurable. A horse's legs and feet are sensitive to blows, sprains, and overwork, especially if the horse is young or is worked on hard surfaces. Lameness may be caused by bony growths, such as splints, spavins, and ringbones, by soft-tissue enlargements, known as windgalls, thoroughpins, and shoe boils, and by injury to the hooves, including sand crack, split hoof, tread thrush, and acute or chronic laminitis.

BREEDS OF HORSES

The first intensively domesticated horses were developed in Central Asia. They were small, lightweight, and stocky. In time, two general groups of horses emerged: the southerly Arab-Barb types (from the Barbary coast) and the northerly, so-called cold-blooded types. When, where, and how these horses appeared is disputed. Nevertheless, all modern breeds—the light, fast, spirited breeds typified by the modern Arabian, the heavier, slower, and calmer working breeds typified by the Belgian, and the intermediate breeds typified by the Thoroughbred—may be classified according to where they originated (*e.g.,* Percheron, Clydesdale, and Arabian), by the principal use of the horse (riding, draft, coach horse), and by their outward appearance and size (light, heavy, pony).

Light horses. *Arabian.* Its long history is obscured by legend, but the Arabian breed, prized for its stamina, intelligence, and character, is known to have been developed in Arabia by the 7th century AD. It is a compact horse with a small head, protruding eyes, wide nostrils, marked withers, and a short back. It usually has only 23 vertebrae, while 24 is the usual number for other breeds. Its legs are strong with fine hooves. The coat, tail, and mane are of fine silky hair. While many colours are possible in the breed, gray prevails. The most famous stud farm is in the region of Najd, Saudi Arabia, but many fine Arabian horses are bred in the United States.

Thoroughbred. The history of the English Thoroughbred is a long one. Records indicate that a stock of Arab and Barb horses was introduced into England as early as the 3rd century. Conditions of climate, soil, and water favoured development, and selective breeding was long encouraged by those interested in racing. Under the reigns of James I and Charles I, 43 mares (the Royal Mares) were imported into England, and a record, the General Stud Book, was begun in which are inscribed only those horses that may be traced back to the Royal Mares in direct line, or to only three other horses imported to England: the Byerly Turk (imported in 1689), the Darley Arabian (after 1700), and the Godolphin Barb (also known as the Godolphin Arabian, imported about 1730). The English Thoroughbred has since been introduced to most countries, where it is bred for racing or used to improve local breeds. The Thoroughbred has a small fine head, a

Thoroughbred lineage

deep chest, and a straight back. Its legs have short bones that allow a long easy stride, and its coat is generally bay or chestnut, rarely black or gray.

Asian. Asian breeds were strongly influenced by Arabian or Persian breeds, which together with the horses of the steppes produced small, plain-looking horses of great intelligence and endurance. Among them are the Tartar, Kirghis, Mongol, and Cossack horses. A Persian stallion and a Dutch mare produced the Orlov trotter in 1778, named after Count Orlov, the owner of the stud farm where the mating took place.

Anglo-Arab. The Anglo-Arab breed originated in France with a crossing of English Thoroughbreds to pure Arabians. The matings produced a horse larger than the Arabian and smaller than the Thoroughbred, of easy maintenance, and capable of carrying considerable weight in the saddle. Its coat is generally chestnut or bay.

American breeds. The Standardbred, a breed that excels at the pace and trot, ranks as one of the world's finest harness racers. A powerful, long-bodied horse, the Standardbred was developed during the first half of the 19th century and can be traced largely to the sire Messenger, a Thoroughbred imported from Britain in 1788 and mated to various brood mares in New York, New Jersey, and Pennsylvania.

The quarter horse was bred for races of a quarter of a mile and is said to descend from Janus, a small Thoroughbred stallion imported into Virginia toward the end of the 18th century. It is 14.2 to 16 hands high, with sturdily muscled hindquarters, essential for the fast departure required in short races. It serves as a polo pony equally well as for ranch work.

The Morgan horse originated from a stallion given to Justin Morgan of Vermont around 1795. This breed has become a most versatile horse for riding, pulling carriages, farm labour, and cattle cutting. It was the ideal army charger. It stands about 15 hands high and is robust, good-natured, willing, and intelligent. Its coat is dark brown or liver chestnut.

The Appaloosa

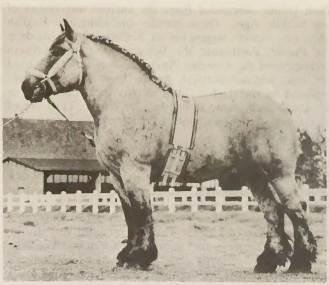
Appaloosa is a colour breed (see above) said to have descended in the Nez Percé Indian territory of North America from wild mustangs, which in turn descended from Spanish horses brought to the New World by explorers. The Appaloosa is 14.2 to 15.2 hands high, of sturdy build and of most diverse use; it is especially good in farm work.

American breeders have also developed several horses that have specialized gaits. These gaited breeds include the American saddle horse, the Tennessee walking horse, and the Missouri fox trotting horse. The American saddle horse has a small head and spectacular high-stepping movements. It is trained for either three or five gaits. The three-gaited horses perform the walk, trot, and canter; the five-gaited horses in addition perform the rack, a quick, high-stepping four-beat gait, and the slow gait, a somewhat slower form of the rack. Since they are used mainly for shows, their hooves are kept rather long, and the muscles of the tail are often clipped so that the base of the tail is carried high. Chestnut and bay are the usual colours. The Tennessee walking horse—a breed derived partially from the Thoroughbred, Standardbred, Morgan, and American saddle horse—serves as a comfortable riding mount used to cover great distances at considerable speed. Its specialty is the running walk, a long and swift stride. Bay is the most common colour. The Missouri fox trotting horse, a breed developed to cover the rough terrain of the Ozark region, is characterized by an unusual gait, called the fox trot, in which the front legs move at a walk while the hind legs perform a trot. The most common colours for this breed are sorrel and chestnut sorrel.

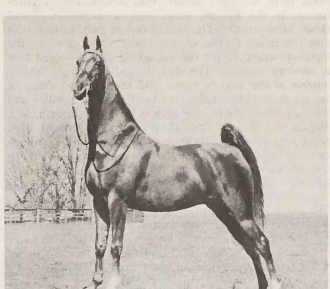
Other light breeds. The English Hackney is a light carriage horse, influenced by the Thoroughbred, capable of covering distances of 12 to 15 miles (19 to 24 kilometres) per hour at the trot and canter. It measures 15.2 to 15.3 hands high and is appreciated for its high knee action.

The Cleveland Bay carriage horse, up to 17 hands high and generally bay in colour, is similar to the Yorkshire Coach horse. Both breeds are now used for the sport of driving.

Other versatile breeds include the German Holstein, Hanoverian, and East Prussian (Trakehner), which serve equally well for riding, light labour, and carriage. These horses, 16 to 18 hands high and of all colours, are now mostly bred for sport.







Representative horse breeds.
(Top) The Belgian, a heavy breed; (centre) an Indian pony; (bottom) the American saddle horse, a light breed.
Sally Anne Thompson

The Andalusian, a high-stepping, spirited horse, and the small but enduring Barb produced the Lipizzaner, which was named after the stud farm founded near Trieste, Italy, in 1580. Originally of all colours, the Lipizzaner is gray or, now exceptionally, bay. It is small, rarely over 15 hands high, of powerful build but slender legs, and with long silky mane and tail. Intelligence and sweetness of disposition as well as gracefulness destined it for academic horsemanship, notably as practiced at the Spanish Riding School of Vienna.

Heavy breeds. The horses used for heavy loads and

farm labour descended from the ancient war horses of the Middle Ages. These breeds—including the English Shire (the world's largest horse), Suffolk, and Clydesdale; the French Percheron; the Belgian horse; the German Noriker; and the Austrian Pinzgauer—are now little used for their original purpose, having been almost entirely replaced by the tractor. They usually measure well over 16 hands high, some more than 19 hands. They are of all colours, sometimes spotted, and generally have a very calm temperament.

Originating in the South Tyrol, the Haflinger is a mountain horse, enduring, robust, and versatile, used for all farm labour, for pulling a carriage or sledge, and for pack hauling. It is rarely over 14.2 hands high and is chestnut with a flaxen mane and tail.

Ponies. Ponies are any horses other than Arabians that are shorter than 14.2 hands. They are generally very sturdy, intelligent, energetic, and sometimes stubborn. The coat is of all colours, mainly dark, and the mane and tail are full. Ponies are used for pulling carriages and pack loads and as children's riding horses or pets. There are numerous varieties, including the Welsh, Dartmoor, Exmoor, Connemara, New Forest, Highland, Dale, Fell, pony of the Americas, Shetland (under seven hands high), Iceland, and Norwegian. Ponies of the warmer countries include the Indian, Java, Manila, and Argentine. (A.W.P./Ed.)

EVOLUTION OF THE HORSE

The evolutionary lineage of the horse is among the best documented in all paleontology. The history of the horse family, Equidae, began during the Eocene Epoch, which lasted from about 54,000,000 to 38,000,000 years ago. During the early Eocene there appeared the first ancestral horse, a hoofed, browsing mammal known technically as *Hyracotherium* but more commonly called eohippus, the "dawn horse." Fossils of eohippus, which have been found in both North America and Europe, show an animal that stood from 4.2 to 5 hands high, diminutive by comparison to the modern horse, with an arched back and raised hindquarters. The legs ended in padded feet with four functional hooves on the forefeet and three on the hindfeet—quite unlike the unpadded, single-hoofed foot of modern equines. The skull lacked the large, flexible muzzle of the modern horse, and the size and shape of the cranium indicate that the brain was far smaller and less complex than that of today's horse. The teeth, too, differed significantly from those of the modern equines, being adapted to a fairly general browser's diet. Eohippus was, in fact, so unhorselike that its evolutionary relationship to the modern equines was at first unsuspected. It was not until paleontologists had unearthed fossils of later extinct horses that the link to eohippus became clear.

Trends in the evolution of the horse

The line leading from eohippus to the modern horse exhibits the following evolutionary trends: increase in size, reduction in the number of hooves, loss of the foot pads, lengthening of the legs, fusion of the independent bones of the lower legs, elongation of the muzzle, increase in the size and complexity of the brain, and the development of crested, high-crowned teeth suited to grazing. This is not to imply that there was a steady, gradual progression in these characteristics leading inevitably from those of eohippus to those of the modern horse. Some of these features, such as grazing dentition, appear abruptly in the fossil record, rather than as the culmination of numerous, gradual changes. Eohippus, moreover, gave rise to many now-extinct branches of the horse family, some of which differed substantially from the line leading to the modern equines.

Although eohippus fossils occur in both the Old and New worlds, the subsequent evolution of the horse took place chiefly in North America. During the remainder of the Eocene, the prime evolutionary changes were in dentition. *Orohippus,* a genus from the middle Eocene, and *Epihippus,* a genus from the late Eocene, resembled eohippus in size and in the structure of the limbs. But the form of the cheek teeth—the four premolars and the three molars found in each half of both jaws—had changed somewhat. In eohippus the premolars and molars were clearly distinct, the molars being larger. In *Orohippus* the fourth premolar had become similar to the molars, and in *Epihippus* both the third and fourth premolars had become molar-like. In addition, the individual cusps that characterized the cheek teeth of eohippus had given way in *Epihippus* to a system of continuous crests or ridges running the length of the molars and molariform premolars. These changes, which represented adaptations to a more specialized browsing diet, were retained by all subsequent ancestors of the modern horse.

Fossils of *Mesohippus,* the next important ancestor of the modern horse, are found in the early and middle Oligocene of North America (the Oligocene Epoch lasted from about 38,000,000 to 26,000,000 years ago). *Mesohippus* was far more horselike than its Eocene ancestors—it was larger (averaging about six hands high); the snout was more muzzlelike; and the legs were longer and more slender. *Mesohippus* also had a larger brain. The fourth toe on the forefoot had been reduced to a vestige, so that both the forefeet and hindfeet carried three functional toes and a foot pad. The teeth remained adapted to browsing.

By the late Oligocene, *Mesohippus* had evolved into a somewhat larger form known as *Miohippus.* The descendants of *Miohippus* split into various evolutionary branches during the early Miocene (the Miocene Epoch lasted from about 26,000,000 to 7,000,000 years ago). One of these branches, known as the anchitheres, included a variety of three-toed, browsing horses comprising several genera. Anchitheres were successful, and some genera spread from North America across the Bering land bridge into Eurasia.

It was a different branch, however, that led from *Miohippus* to the modern horse. The first representative of this line, *Parahippus,* appeared in the early Miocene. *Parahippus* and its descendants marked a radical departure in that they had teeth adapted to eating grass. Grasses were at this time becoming widespread across the North American plains, providing *Parahippus* with a vast food supply. Grass is a much coarser food than succulent leaves and requires a different kind of tooth structure. The cheek teeth developed larger, stronger crests and became adapted to the side-to-side motion of the lower jaw necessary to grind grass blades. Each tooth also had an extremely long crown, most of which, in the young animal, was buried beneath the gum line. As grinding wore down the exposed surface, some of the buried crown grew out. This high-crowned tooth structure assured the animal of having an adequate grinding surface throughout its normal life span. Adaptations in the digestive tract must have occurred as well, but the organs of digestion are not preserved in the fossil record.

Dental adaptations to a grass diet

The change from browsing to grazing dentition was essentially completed in *Merychippus,* which evolved from *Parahippus* during the middle and late Miocene. *Merychippus* must have looked much like a modern pony. It was fairly large, standing about 10 hands high, and its skull was similar to that of the modern horse. The long bones of the lower leg had become fused; this structure, which has been preserved in all modern equines, is an adaptation for swift running. The feet remained three-toed, but in many species the foot pad was lost, and the two side toes became rather small. In these forms, the large central toe bore the animal's weight. Strong ligaments attached this hoofed central toe to the bones of the ankles and lower leg, providing a spring mechanism that pushed the flexed hoof forward after the impact of hitting the ground. *Merychippus* gave rise to numerous evolutionary lines during the late Miocene. Most of these, including *Hipparion, Neohipparion,* and *Nannippus,* retained the three-toed foot of their ancestors. One line, however, led to the one-toed *Pliohippus,* the direct predecessor of *Equus. Pliohippus* fossils occur in the early to middle Pliocene beds of North America (the Pliocene Epoch lasted from about 7,000,000 to 2,500,000 years ago).

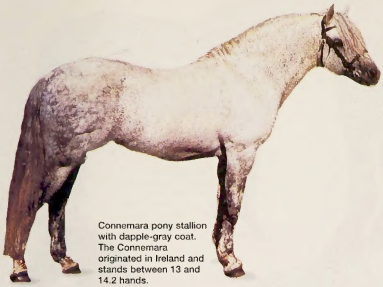
Equus—the genus to which all modern equines, including horses, asses, and zebras, belong—evolved from *Pliohippus* toward the end of the Pliocene. *Equus* shows even greater development of the spring mechanism in the foot and exhibits straighter and longer cheek teeth. This new form was extremely successful and had spread from



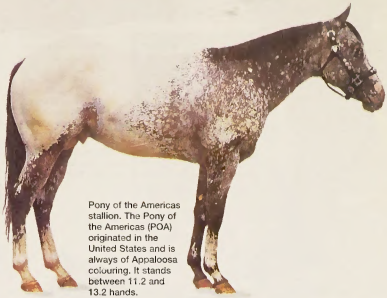
Shetland pony stallion with chocolate-coloured coat and flaxen mane and tail. It originates in Scotland, and its average height is 40 inches.



Welsh pony stallion with white coat. The Welsh pony originated in Wales and stands between 12.2 and 13.2 hands.



Connemara pony stallion with dapple-gray coat. The Connemara originated in Ireland and stands between 13 and 14.2 hands.



Pony of the Americas stallion. The Pony of the Americas (POA) originated in the United States and is always of Appaloosa colouring. It stands between 11.2 and 13.2 hands.

(Immediate right) A young girl guides her Welsh pony over a jump during competition.

(Far right) An Icelandic horse moving swiftly at the *tölt*, a smooth four-beat, lateral running walk. Icelandics carry adults with ease. Although this breed shares its small stature with ponies, standing between 12.3 and 13.2 hands, Icelanders refer to it as a horse.

(Bottom right) A team of Haflingers at a driving demonstration. The Haflinger is from the mountains of Austria's Tyrol. Its coat is palomino or a rich chestnut with a flaxen mane and tail. Average height is 13.3 hands.







Plate 1: (Top left and right, upper middle right) Scott Smockley—EB Inc. Upper middle left © Sally Anne Thompson Animal Photography. Bottom left and right © Bob Langrish. Lower middle right © Pal Stefansson/Iceland Review

Light Horses



American Quarter Horse stallion with buckskin coat. The Quarter Horse originated in the United States and averages between 14.2 and 16 hands.



Appaloosa mare with bay colouring. This spotted breed descends from the prized horses of the Nez Percé of the northwestern United States. It averages between 14.2 and 16 hands.



Morgan stallion with bay coat. The Morgan originated in the United States and averages between 14.1 and 15.2 hands.



American Paint Horse mare of bay colouring. This colour breed originated in the United States and averages between 15 and 16 hands.



Arabian gelding with chestnut coat. This ancient breed originated in the Middle East and averages between 14 and 15 hands.



Standardbred gelding with dark bay coat. The Standardbred originated in the United States and averages between 15 and 16 hands.

An Arabian is guided up a steep mountain slope during an endurance race. The Arabian, with its speed and stamina, dominates the sport of endurance racing.





Hanoverian stallion with dark bay coat. The Hanoverian originated in Germany and averages between 15.3 and 17 hands.

A Dutch Warmblood stallion negotiates a fence during a show jumping competition. The Dutch Warmblood originated in The Netherlands and has an average height of 16.2 hands.





Trakehner stallion with dark bay coat. The Trakehner originated in a part of Russia that was once East Prussia. It averages between 16 and 17 hands.

An American Quarter Horse is brought to a sliding stop in a reining competition. The sport of reining tests the horse's athletic ability and training in a series of maneuvers called a "reining pattern," where the horse performs spins, circles, and sliding stops, often with great speed.





Cleveland Bay stallion. The Cleveland Bay originated in England and is always of bay colouring. It averages between 16 and 16.2 hands.

A Hanoverian is guided at the canter during a dressage test.



Plate 2: All photographs, Scott Smuckler—EB Inc.

Plate 3: (Top left, upper middle left, lower middle left) Scott Smuckler—EB Inc.; (bottom left) © Bob Langrish; (top right) © Melinda Hughes Photography; (upper middle right) © Tish Quirk; (lower middle right) © Don Trout; (bottom right) © Karl Leck/USESA



Thoroughbred stallion with dark bay coat. The Thoroughbred originated in England and averages between 15 and 17 hands.



Missouri Fox Trotting mare with black coat. The Missouri Fox Trotter originated in the United States and averages between 14 and 16 hands.



Lipizzaner stallion with white coat. The Lipizzaner originated in a part of Slovenia that was once ruled by Austria. It averages between 15 and 16.1 hands.



Tennessee Walking Horse stallion with dapple-gray coat. The Tennessee Walker originated in the United States and averages between 15 and 16 hands.



Paso Fino gelding with palomino coat. The Paso Fino originated in Puerto Rico and averages between 14 and 15 hands.



Akhal-Teke stallion with golden dun coat. The Akhal-Teke originated in Turkmenistan and averages between 14.2 and 16 hands.



Argentine Criollo with grulla coat. The Criollo originated in Argentina and has an average height of 14 hands.



American Saddlebred mare with black coat. The American Saddlebred originated in the United States and averages between 15 and 16 hands.



Andalusian stallion with dark gray coat. The Andalusian originated in Spain and averages between 15.1 and 15.3 hands.











(Left) An American Saddlebred performing an animated trot. (Top centre) A Peruvian Paso, wearing traditional tack from Peru, executing its smooth riding gait. (Bottom centre) A Thoroughbred leaps an obstacle during the cross-country phase of a three-day event. (Top right) A Hackney horse performs its typical high-stepping trot during a driving competition. The Hackney originated in England and averages between 14 and 15.3 hands. (Bottom right) A palomino American Quarter Horse cutting a cow from a herd.

Plate 4: All photographs, Scott Smudsky—EB Inc.

Plate 5: (Top left and right, lower middle right) © Sally Anne Thompson/Animal Photography, (upper middle left and right) Scott Smudsky—EB Inc., (lower middle left) © Sue Lewis, (lower middle centre) © Jackie Jacobson, (bottom left) © Gemma Giannini, (bottom right) © Bob Langrish

Heavy Horses



Belgian stallion with sorrel coat and flaxen mane and tail. The Belgian originated in Belgium and averages between 15.3 and 17 hands.



Shire stallion with bay coat. The Shire originated in England and is the world's largest horse, with an average height of 17 hands (sometimes reaching 19 hands).



Clydesdale stallion with bay coat. The Clydesdale originated in Scotland and averages between 16.1 and 18 hands.



Percheron stallion with black coat. The Percheron originated in France and has an average height of 16 hands.

A team of Clydesdales pulling a plow at a draft-horse demonstration. The horse on the left has a strawberry, or red roan, coat; the horse on the right has a blue roan coat.



the plains of North America to South America and to all parts of the Old World by the early Pleistocene (the Pleistocene Epoch lasted from about 2,500,000 to 10,000 years ago). *Equus* flourished in its North American homeland throughout the Pleistocene; then, about 10,000 to 8,000 years ago, *Equus* disappeared from North and South America. Scholars have offered various explanations for this disappearance, including the emergence of devastating diseases or the arrival of human populations (which presumably hunted the horse for food). Despite these speculations, the reasons for the demise of *Equus* in the New World remain uncertain. The submergence of the Bering land bridge prevented any return migration of horses from Asia, and *Equus* was not reintroduced into its native continent until the Spanish explorers brought horses in the early 16th century.

The demise of *Equus* in the Americas

During the Pleistocene, the evolution of *Equus* in the Old World gave rise to all the modern members of the genus. The modern horse, *Equus caballus*, became widespread from central Asia to most of Europe. Local types of horses, all breeds of this single species, undoubtedly developed, and three of these—Przewalski's horse from central Asia, the tarpan from eastern Europe and the Ukrainian steppes, and the forest horse of northern Europe—are generally credited as being the ancestral stock of the domestic horse. According to this line of thinking, Przewalski's horse and the tarpan formed the basic breeding stock from which the southerly "warm-blooded" horses developed, while the forest horse gave rise to the heavy, "cold-blooded" breeds.

(Ed.)

Riding and horsemanship

HISTORY

Origins and early history. From the 2nd millennium BC, and probably even earlier, the horse was employed as a riding animal by fierce nomadic peoples of central Asia. One of these peoples, the Scythians, were accomplished horsemen and used saddles. It is also likely that they realized the importance of a firm seat and were the first to devise a form of stirrup. A saddled horse with straps hanging at the side and looped at the lower end is portrayed on a vase of the 4th century BC found at Chertomlyk in the Soviet Union. This contrivance may have been used for mounting only, however, because of the danger of being unable to free the foot quickly in dismounting. The Greek historian Strabo said that the indocility of the Scythians' wild horses made gelding necessary, a practice until then unknown in the ancient world. The Sarmatians, superb horsemen who superseded the Scythians, rode bareback, controlling their horses with knee pressure and distribution of the rider's weight.

Among the earliest peoples to fight and hunt on horseback were the Hittites, the Assyrians, and the Babylonians; at the same time (about 1500 BC) the Hyksos, or Shepherd Kings, introduced horses into Egypt and rode them in all their wars. In the 8th and 7th centuries BC, the Scythians brought horses to Greece, where the art of riding developed rapidly, at first only for pleasure. A frieze from the Parthenon in Athens shows Greeks riding bareback. Philip II of Macedon had a body of cavalry in his army, and the army of his son Alexander had separate, organized horse units. In the 4th century BC another Greek historian, Xenophon, wrote his treatise *Peri hippikēs* (*On Horsemanship*), giving excellent advice on horsemanship. Many of his principles are still perfectly valid. He advocated the use of the mildest possible bits and disapproved of the use of force in training and in riding. The Roman mounted troops were normally barbarian archers who rode without stirrups and apparently without reins, leaving the hands free to use the bow and arrow.

Classical horsemanship

As a general rule almost every item of riding equipment used today originated among the horsemen of the Eurasian steppes and was adopted by the people of the lands they overran to the east, the south, and later the west.

Horseshoes of various types were used by migratory Eurasian tribes about the 2nd century BC, but the nailed iron horseshoe as used today first appeared in Europe about the 5th century AD, introduced by invaders from the East. One, complete with nails, was found in the tomb of the Frankish king Childeric I at Tournai, Belg.

Attila is said to have brought the stirrup to Europe. Round or triangular iron stirrups were used by the Avars in the 6th century AD, and metal stirrups were used by the Byzantine cavalry. They were in use in China and Japan by about AD 600.

The principle of controlling a horse by exerting pressure on its mouth through a bit (a metal contrivance inserted in the mouth of the horse) and reins (straps attached to the bit held by the rider) was practiced from the earliest times, and bits made of bone and antlers have been found dating from before 1000 BC. The flexible mouthpiece with two links and its variations have been in use down the centuries, leading directly to the jointed snaffle bit of the present day.

Early, stumpy prickspurs have been found in Bohemia on 4th-century-BC Celtic sites.

Military horsemanship. The importance of cavalry increased in the early Middle Ages, and in the 1,000 years that followed, mounted warriors became predominant in battle. Armour steadily became bulkier and heavier, forcing the breeding of more and more massive horses, until the combination rendered maneuverability nearly impossible.

Efforts to overcome this were made at a Naples riding academy in the early 16th century, when Federico Grisone and Giovanni Battista Pignatelli tried to combine classical Greek principles with the requirements of medieval mounted combat. After Xenophon, except for a 14th-century treatise by Ibn Hudhayl, an Arab of Granada, Spain, apparently no literature on riding was produced until Grisone published his *Gli ordini di cavalcare* ("The Orders of Riding") in 1550.

The development of firearms led to the shedding of armour, making it possible for some further modifications in methods and training under followers of the school of Pignatelli and Grisone, such as William Cavendish, duke of Newcastle. In 1733 François Robichon de la Guérinière published *École de cavalerie* ("School of Cavalry"), in which he explained how a horse can be trained without being forced into submission, the fundamental precept of modern dressage. Dressage is the methodical training of a horse for any of a wide range of purposes, excluding only racing and cross-country riding.

Meanwhile, the Spanish Imperial Riding School in Vienna and the French cavalry centre at Saumur aimed at perfecting the combined performance of horse and rider. Their technique and academic seat, a formal riding position or style in which the rider sits erectly, deep in the middle of the saddle, exerted considerable influence in Europe and America during the 18th and 19th centuries and are still used in modern dressage. The head riding master at Saumur, Comte Antoine d'Aure, however, promoted a bold, relaxed, and more natural, if less "correct," style of riding across country, in disagreement with his 19th-century contemporary François Baucher, a horseman of great ability with formal haute école ("high school") ideas. Classical exercises in the manège, or school for riding, had to make way for simplified and more rational riding in war and the hunt. During this period hunting riders jumped obstacles with their feet forward, their torso back on the horse's haunches, and the horse's head held up. The horse often leaped in terror.

Growth of modern technique

At the turn of the 20th century, Capt. Federico Caprilli, an Italian cavalry instructor, made a thorough study of the psychology and mechanics of locomotion of the horse. He completely revolutionized the established system by innovating the forward seat, a position and style of riding in which the rider's weight is centred forward in the saddle, over the horse's withers. Caprilli wrote very little, but his pupil, Piero Santini, popularized his master's fundamental principles. Except in dressage and showing, the forward seat is the one now most frequently used, especially for jumping.

THE ART OF HORSEMANSHIP

The basic principle of horsemanship is to obtain results in a humane way by a combination of balance, seat, hands, and legs.

Fundamentals. The horse's natural centre of gravity shifts with its every movement and change of gait. Considering that a mounted horse also carries a comparatively unstable burden approximately one-fifth of its own weight, it is up to the rider to conform with the movements of the horse as much as possible.

Before one mounts, the saddle is checked to be sure that it fits both the horse and its rider. Experienced riders position themselves in the saddle in such a way as to be able to stay on the horse and control it. The seat adopted depends on the particular task at hand. A secure seat is essential, giving riders complete independence and freedom to apply effectively the aids at their disposal. Good riders do not overrule the horse, but, firmly and without inflicting pain, they persuade it to submit to their wishes.

The horse's movements. The natural gaits of the horse are the walk, the trot, the canter or slow gallop, and the gallop, although in dressage the canter and gallop are not usually differentiated. A riding horse is trained in each gait and in the change from one to another.

During the walk and the gallop the horse's head moves down and forward, then up and back (only at the trot is it still); riders follow these movements with their hands.

Walk. The walk is a slow, four-beat, rhythmic pace of distinct successive hoof beats in an order such as near (left) hind, near fore, off (right) hind, off fore. Alternately two or three feet may be touching the ground simultaneously. It may be a free, or ordinary, walk in which relaxed extended action allows the horse freedom of its head and neck, but contact with the mouth is maintained; or it may be a collected walk, a short-striding gait full of impulsion, or vigour; or it may be an extended walk of long, unhurried strides.

Trot. The trot is a two-beat gait, light and balanced, the fore and hind diagonal pairs of legs following each other almost simultaneously—near fore, off hind, off fore, and near hind. Riders can either sit in the saddle and be bumped as the horse springs from one diagonal to the other, or they can rise to the trot, post, by rising out of the saddle slightly and allowing more of their weight to bear on the stirrups when one or the other of the diagonal pairs of legs leaves the ground. Posting reduces the impact of the trot on both horse and rider.

Canter. As the horse moves faster, its gait changes into the canter, or ordinary gallop, in which the rider does not rise or bump. It is a three-beat gait, graceful and elegant, characterized by one or the other of the forelegs and both hindlegs leading—near hind, off hind, and near fore practically together, then off fore, followed briefly by complete suspension. Cantering can be on the near lead or the off, depending on which is the last foot to leave the ground. The rider's body is more forward than at the trot, the weight taken by the stirrups.

Gallop. An accelerated canter becomes the gallop, in which the rider's weight is brought sharply forward as the horse reaches speeds up to 30 miles (48 kilometres) an hour. The horse's movements are the same as in the canter. To some authorities, the gallop is a four-beat gait, especially in an extended run.

Other gaits. There are a number of disconnected and intermediate gaits, some done only by horses bred to perform them. One is the rack, a four-beat gait, with each beat evenly spaced in perfect cadence and rapid succession. The legs on either side move together, the hindleg striking the ground slightly before the foreleg. The single foot is similar to the rack but slower. In the pace, the legs on either side move and strike the ground together in a two-beat gait. The fox trot and the amble are four-beat gaits, the latter smoother and gliding.

The rack gait

Training and equipment. Depending on the abilities and inclinations of horse and trainer, training may include such elements as collection (controlled, precise, elevated movement) and extension (smooth, swift, reaching movement—the opposite of collection) at all paces; turns on the forehand (that part of the horse that is in front of the rider) and hindquarters; changing lead leg at the canter; change of speed; reining back, or moving backward; lateral movements; and finally the refinements of dressage, jumping, and cross-country riding.

Communication with the horse is rendered possible by the use of the bit and the aids. The rider signals intentions to the horse by a combination of recognized movements of hands and legs, using several articles of equipment. By repetition the horse remembers this language, understands what is required, and obeys.

Bits. There are several types of bits, including the snaffle, the double bridle, and the Pelham.

The simplest is the snaffle, also called the bridoon. It consists of a single straight or jointed mouthpiece with a ring at each end for the reins. The snaffle is used for racing and frequently for cross-country riding. It is appropriate for preliminary schooling.

The double bridle is used for advanced schooling. It consists of a jointed snaffle and a straight bit placed together in the mouth, first the snaffle, then the bit, both functioning independently and attached to separate reins. The mouthpiece of the bit can have a port or indentation in its centre to give more control. The slightest pull on the bit rein exerts pressure on the mouth.

The double bridle

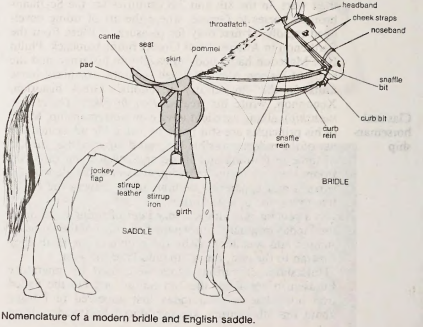
The Pelham is a snaffle with a straight mouthpiece; cheekpieces with rings at the lower ends for curb action; and a curb chain, with which pressure may be applied to the lower outside of the mouth. The Pelham gives control with only slight discomfort and is popular for polo.

Bridles. The bridle is a set of straps that makes the bit secure in the animal's mouth and thus ensures human control by means of the reins (see figure). The upper portion of the bridle consists of the headpiece passing behind the ears and joining the headband over the forehead; the cheek straps run down the sides of the head to the bit, to which they are fastened; in the blind type of driving bridle the blinkers, rectangular or round leather flaps that prevent the animal from seeing anything except what lies in front, are attached to the cheek straps; the noseband passes around the front of the nose just above the nostrils; and the throatlatch extends from the top of the cheek straps underneath the head.

Aids. The principal features of a horse's mentality are acute powers of observation, innate timidity, and a good memory. To a certain extent the horse can also understand. Schooling is based on these faculties, and the rider's aids are applied accordingly. The natural aids are the voice, the hands through the reins and the bit, the legs and heels, and the movement of the rider's weight. The whip, the spur, and devices such as martingales, special nosebands, and reins are artificial aids, so termed in theory, as the horse does not discriminate between natural and artificial.

Horses are easily startled. A good horseman will approach them quietly, speaking to them and patting them to give them confidence. Silence on the part of the rider can even cause disquiet to some horses, but they should not be shouted at. The rider's voice and its tone make

By courtesy of the American Saddle Horse Breeders Assoc. and W.W. Rouch & Co.



Nomenclature of a modern bridle and English saddle.

a useful aid in teaching a horse in its early schooling to walk, trot, canter, and halt.

To keep the horse alert at all times, the rider's hands keep a light, continual contact with its mouth, even at the halt. The hands are employed together with the legs to maintain contact, to urge the horse forward, to turn, to rein back, and generally to control the forehand. The horse is said to be collected and light in hand when the action of the bit can cause it to flex, or relax, its jaw with its head bent at the poll, or top.

When pressed simultaneously against the flanks, immediately after the hands ease the reins, the legs induce the forward movement of the horse. They are of the greatest importance in creating and maintaining impulsion, in controlling the hindquarters, and for lateral movement.

Riders achieve unity of balance by means of the weight aid, that is, by moving the body in harmony with the movements of the horse, forward, backward, or to the side. Thus, in cantering to the left, the rider leans to the left; or when about to descend a steep slope, the rider stays erect while the horse is feeling for the edge with its forefeet, but as soon as the descent starts the rider leans forward, leaving the hindquarters free to act as a brake and to prevent scraping the back of the horse's rear legs on rough ground. Meanwhile the hands keep the horse headed straight to maintain its balance.

The use of the whip

The whip is used chiefly to reinforce the leg aid for control, to command attention, and to demand obedience, but it can be used as a punishment in cases of deliberate rebellion. A horse may show resistance by gnashing its teeth and swishing its tail. Striking should always be on the quarters, behind the saddle girth, and must be immediate since a horse can associate only nearly simultaneous events. This applies equally to rewards. A friendly tone of voice or a pat on the neck are types of reward.

Although normally the leg or the heel, or both, should be sufficient, spurs, which should always be blunt, assist the legs in directing the precision movements of advanced schooling. Their use must be correctly timed.

Martingales are of three types: running, standing, or Irish. The running and standing martingales are attached to the saddle straps at one end and the bit reins or bridle at the other. The Irish martingale, a short strap below the horse's chin through which the reins pass, is used for racing and stops the horse from jerking the reins over its head. As the horse cannot see below a line from the eye to the nostril, it should not be allowed to toss its head back, particularly near an obstacle, as it is liable to leap blindly. A martingale should not be necessary with a well-schooled horse.

The noseband, a strap of the bridle that encircles the horse's nose, may be either a cavesson, with a headpiece and rings for attaching a long training rein, or a noseband with a headstrap, only necessary if a standing martingale is used. A variety of other nosebands are intended for horses that pull, or bear, on the reins unnecessarily.

Seats. The saddle, the length of the stirrup, and the rider's seat, or style of riding, should suit the purpose for which the horse is ridden. The first use of the stirrup is to enable the rider to get on the horse, normally from the near (left) side. With the raised foot in the stirrup the rider should avoid digging the horse in the flank on springing up and should gradually slide into position without landing on the horse's kidneys with a bump. With an excitable horse, the rider may wait, resting on knees and stirrups, until the horse moves forward.

Forward seat. The forward seat, favoured for show jumping, hunting, and cross-country riding, is generally considered to conform with the natural action of the horse. The rider sits near the middle of the saddle, his torso a trifle forward, even at the halt. The saddle is shaped with the flaps forward, sometimes with knee rolls for added support in jumping. The length of the stirrup leather is such that, with continual lower thigh and knee grip, the arch of the foot can press on the tread of the iron with the heel well down. A wide and heavy stirrup iron allows easy release of the foot in case of accidents. The line along the forearm from the elbow to the hands and along the reins to the bit is held straight. As the horse moves forward, so do the rider's hands, to suit the horse's comfort.

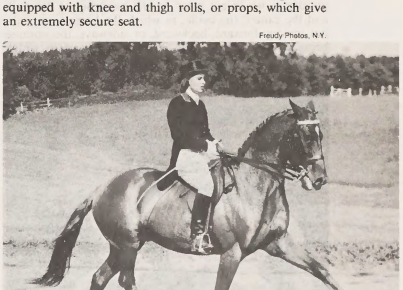


Forward seat in jumping.
By courtesy of the American Horse Shows Association

Dressage seat. In the show and dressage seat the rider sinks deep into the saddle, in a supple, relaxed but erect position above it. The saddle flaps are practically straight so as to show as much expanse of the horse's front as possible. The stirrup leather is of sufficient length for the rider's knee to bend at an angle of about 140 degrees and for the calf to make light contact with the horse's flank, the heel well down, and the toes or the ball of the foot resting on the tread of the stirrup iron. The rider keeps continual, light contact with the horse's mouth; and the intention is to convey an impression of graceful, collected action. In the past this type of saddle, with its straight-cut flaps, was used for hunting and polo, but the forward seat has become more popular for these activities.

Stock saddle. The stock saddle seat is appropriate for ranchers but is also used at rodeos and by many pleasure and trail riders. The saddle, which can weigh as much as 40 pounds (18 kilograms), is designed for rounding up cattle and is distinguished by a high pommel horn for tying a lariat. The rider employs long stirrups and a severe bit that he seldom uses because he rides with a loose rein, guiding his horse chiefly by means of shifting the weight of his body in the saddle. The gaucho roughriders of the Argentine Pampa have adopted a similar seat, using a saddle with a high pommel and cantle. Australian stockmen have used a saddle that has a short flap and is equipped with knee and thigh rolls, or props, which give an extremely secure seat.

Freudy Photos, N.Y.



Dressage seat in the extended trot.

Side saddle. Though now not so fashionable, the elegant and classical side-saddle seat was formerly favoured and considered correct by many horsewomen. On the near side the saddle has an upright pommel on which the rider's right leg rests. There is a lower, or leaping, pommel, against which the left leg can push upward when grip is required, and a single stirrup. Although the rider sits with both legs on one side of the saddle, forward action to suit the movement of the horse is feasible in cross-country riding.

Bareback. Bareback means riding without saddle or blanket, the rider sitting in the hollow of the horse's back and staying there chiefly by balance. It is an uncomfortable seat but less so at the walk and the slow canter. When suffering from saddle galls horses are sometimes ridden bareback for exercise.

Dressage. Originally intended for military use, dressage training was begun early in the 16th century. The international rules for dressage are based on the traditions and practice of the best riding schools in the world. The following is an extract from these rules of the Fédération Équestre Internationale:

Object and general principles.

The object of dressage is the harmonious development of the physique and the ability of the horse. As a result, it makes the horse calm, supple, and keen, thus achieving perfect understanding with its rider. These qualities are revealed by the freedom and regularity of the paces; the harmony, lightness, and ease of the movements; the lightening of the forehand, and the engagement of the hindquarters; the horse remaining absolutely straight in any movement along a straight line, and bending accordingly when moving on curved lines.

The horse thus gives the impression of doing of his own account what is required of him. Confident and attentive, he submits generously to the control of his rider. (Used with permission of the publisher.)

Campagne is the term used for elementary but thorough training, including work on the longeing rein. This long rein, also used for training young or difficult horses, is attached to a headpiece with a noseband called a cavesson. The horse is bitted and saddled and is schooled in circles at the end of the rein. It is an accessory to training from the saddle, which is always best. Basic to campagne is collection: teaching the horse to arch its neck, to shift its weight backward onto its hindquarters, and to move in a showy, animated manner. Other elements of campagne include riding in a straight line, turns, and lateral movements.

Haute école dressage

Haute école is the most elaborate and specialized form of dressage, reaching its ultimate development at the Vienna school in its traditional white Lippizaner horses. Some characteristic haute école airs, or movements, are the pirouettes, which are turns on the haunches at the walk and the canter; the piaffe, in which the horse trots without moving forward, backward, or sideways, the impulse being upward; the passage, high-stepping trot in which the impulse is more upward than forward; the levade, in which the horse stands balanced on its hindlegs, its forelegs drawn in; the courvet, which is a jump forward in the levade position; and the croupade, ballotade, and capriole, a variety of spectacular airs in which the horse jumps and lands again in the same spot.

All of these movements are based, perhaps remotely in some instances, on those that the horse performs naturally.

Jumping. The most sensitive parts of the horse when ridden are the mouth and the loins, particularly in jumping. The rider's hands control the forehand while the legs act on the hindquarters. As speed is increased the seat is raised slightly from the saddle, with the back straight and the trunk and hands forward, the lower thighs and the knees taking the weight of the body and gripping the saddle, leaving the legs from the knees down free for impulsion. Contact with the mouth is maintained evenly and continually, the rider conforming with every movement as the horse's head goes forward after takeoff and as it is retracted on landing, the hands always moving in line with the horse's shoulder. In order to give complete freedom to the hindquarters and to the hocks, the rider does not sit back in the saddle until at least two strides after landing.

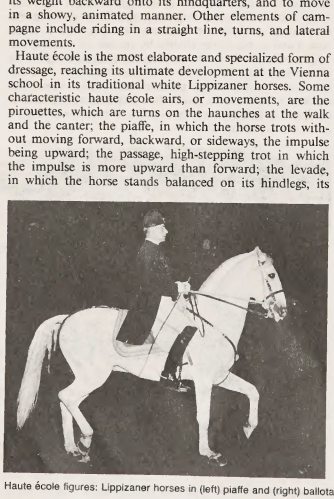
The horse is a natural jumper, but, if ridden, schooling becomes necessary. Training is started in an enclosed level area by walking the horse, preferably in a snaffle, over a number of bars or poles laid flat on the ground. When the horse has become accustomed to this, its speed is increased. As the horse progresses, the obstacles are systematically raised, varied, and spaced irregularly. The object is to teach the horse: (1) to keep its head down; (2) to approach an obstacle at a quiet, collected, yet energetic pace; (3) to decide how and where to take off; and (4) after landing to proceed quietly to the next obstacle. The horse should be confident over every jump before it is raised and should be familiarized with a variety of obstacles.

Only thoroughly trained riders and horses compete. Very strenuous effort is required of the horse, as well as of the rider who does not by any action give the horse the impression that something out of the ordinary is impending. If possible the horse is warmed up by at least a half-hour's walking and trotting before entering the ring. The horse is guided toward the exact centre of every obstacle, the rider looking straight ahead and not looking around after takeoff for any reason, as that might unbalance the horse. The broader the obstacle, the greater the speed of approach. Although a few experienced riders can adjust the horse's stride for a correct takeoff, this should not be necessary with a well-schooled horse. The rider is always made to conform with every action of the horse, the only assistance necessary being that of direction and increasing or decreasing speed according to the obstacle.

RIDING AND SHOWS

Racing on horseback probably originated soon after man first mastered the horse. By the 7th century BC organized mounted games were held at Olympia. The Romans held race meetings, and in medieval Europe tournaments, jousting, and horse fairs were frequent and popular events.

Freudy Photos





Haute école figures: Lippizaner horses in (left) piaffe and (right) ballotade.

Played in Persia for centuries, polo was brought to England from India about 1870. In North America, Western ranch riding produced the rodeo.

Horse associations and pony clubs are today the mainstay of equine sport. They have improved the standards of riding instruction and the competitive activities of dressage, hunter trials, and show jumping. The latter became an important event from 1869, when what was probably the first "competition for leaping horses" was included in the program of an Agricultural Hall Society horse show in London. National organizations such as the British Horse Society, the American Horse Shows Association (AHSA), the Federazione Italiana Sports Equestri, the National Equestrian Federation of Ireland, the Fédération Française des Sports Équestres, and similar groups from about 50 other nations are affiliated with the Fédération Équestre Internationale (FEI), founded in 1921 with headquarters at Brussels, the official international governing body and the authority on the requirements of equitation.

Associations and clubs

Horse shows. Horse shows are a popular institution that evolved from the horse sections of agricultural fairs. Originally they were informal displays intended to attract buyers and encourage the improvement of every type of horse. Now they are organized and conducted by committees of experts and by associations that enforce uniform rules, appoint judges, settle disputes, maintain records, and disseminate information. Riding contests included in the program have become increasingly important.

Under the auspices of the Royal Dublin Society, an international horse show was first held at Dublin in 1864. It is an annual exhibition of every type of saddle horse, as well as broodmares and ponies. International jumping contests similar to Olympic competition, events for children, and auction sales are held during this five-day show.

The National Horse Show at New York, first held in 1883, is another great yearly event. Held at Madison Square Garden, it lasts several days and includes about 10 different events. Among the most important are the international jumping under FEI rules and the open jumping under AHSA rules. Other shows are held in many sections of the United States.

Horse and pony shows are held regularly in the United Kingdom, the most important being the Richmond Royal Horse Show, the Horse of the Year Show, and the Royal International Horse Show. The latter, an annual event first held in 1907, has flourished under royal patronage and includes international jumping, special items such as the visit of the Spanish Riding School with its Lippizaners in 1953, and a Supreme Riding Horse competition.

In Canada, the Royal Agricultural Winter Fair at Toronto, opened in 1922 and known in Canada as the "Royal," is a major event, and in Australia the Royal Agricultural Society organizes horse shows annually in every state. Other events include the shows at Verona and at the Piazza di Siena in Rome; frequent horse shows in Belgium, France, Germany, and The Netherlands; the winter show in July in Buenos Aires; and the Exhibition of Economic Achievement in Moscow.

Olympic equestrian competition. The FEI organizes and controls the equestrian events at the Olympic Games. Included in each Olympics since the Games at Stockholm in 1912 (equestrian events were also held in 1900), these events are the occasion for keen rivalry and evoke high standards of horsemanship. They comprise a dressage grand prix, a three-day event, and a jumping grand prix, all open to team and individual competition.

The Grand Prix de Dressage involves performance of the walk, trot, canter, and collected paces and several conventional dressage figures and movements, as well as the correct rider's position. Scoring on each item is from a maximum of 10 for excellent down to 1 for very bad.

The three-day event consists of tests in dressage, endurance or cross-country riding, and show jumping. Dressage is on the first day. On the second day there is an endurance test over a course 25 to 35 kilometres (16 to 22 miles) in length, covering swamp roads, tracks, steeplechase obstacles, and cross-country sections. Jumping tests, less strenuous than the Prix des Nations jumping event, are held on the third day.

The Prix des Nations jumping event is a competition involving 13 or 14 obstacles, heights varying between 1.30 and 1.60 metres (51 and 63 inches), and a water jump 4 metres (13 feet) across, over a course with 60 metres (200 feet) between obstacles. Penalties are scored for disobedience, knocking down or touching an obstacle, and for a fall. The rider with the lowest penalty score wins.

In addition to these competitions there is a riding section of the modern pentathlon, also conducted under FEI rules. Competitors must clear, riding a strange horse chosen by lot, 20 obstacles over a course 1,000 metres (3,000 feet) in length. Other international competitions began in the 1950s under the supervision of the FEI. (C.E.C.)

BIBLIOGRAPHY

General works: GEORGE G. SIMPSON, *Horses* (1951, reprinted 1970), a very readable and popular account of the horse family today and through 60,000,000 years of development, with a good bibliography; MARGARET C. SELF, *The Horseman's Encyclopedia*, rev. ed. (1963, reprinted 1978), an invaluable collection of information on domestic horses. Breed associations issue pamphlets on selected breeds of horses. See also *A Standard Guide to Horse and Pony Breeds*, general ed. ELWYN HARTLEY EDWARDS (1980), providing information on 150 breeds and types. *The Horseman's International Book of Reference*, ed. by JEAN FROISSARD and LILY POWELL FROISSARD (1980), is an authoritative source consisting of contributions of international experts; C.E.G. HOPE and G.N. JACKSON (eds.), *The Encyclopedia of the Horse* (1973), is a comprehensive reference work discussing, among other specific topics, the horse in mythology, literature, and art. See also JOHN BASKETT, *The Horse in Art* (1980). Special aspects are examined in DAVID P. WILLOUGHBY, *Growth and Nutrition in the Horse* (1975); R.H. SMYTHE, *The Horse: Structure and Movement*, 2nd ed., rev. by PETER C. GOODY (1972); GEORGE H. WARING, *Horse Behavior: The Behavioral Traits and Adaptations of Domestic and Wild Horses, Including Ponies* (1983); MOYRA WILLIAMS, *Horse Psychology*, rev. ed. (1976); and PETER WILLETT, *The Thoroughbred* (1970). R.S. SUMMERHAYS, *Encyclopaedia for Horsemen*, 6th ed., rev. by STELLA A. WALKER (1975), is also a useful reference work.

Horsemanship: MARY GORDON-WATSON, *The Handbook of Riding* (1982); ALBERT E. DECARPENTRY, *Academic Equitation*, trans. from the French (1971); VLADIMIR S. LITTAUER, *Common Sense Horsemanship*, 2nd ed. (1963, reprinted 1974); HENRY WYNMALEN, *Equitation*, 2nd ed. (1952, reissued 1971); EARL R. FARSCHLER, *Riding and Training*, new ed. (1959, reissued 1972), which contains a description of the gaits; JEAN S.-F. PAILLARD, *Understanding Equitation*, trans. from the French (1974); C.E.G. HOPE, *The Horseman's Manual* (1972); JANE KIDD, *Horsemanship in Europe* (1977); and MYRON J. SMITH, *Equestrian Studies* (1981), a classified bibliography of more than 4,600 English-language items appearing between 1950 and 1980. See also JACKIE SPAULDING, *The Family Horse: How to Choose, Care for, Train and Work Your Horse* (1982). (*History*): STAN STEINER, *Dark and Dashing Horsemen* (1981); CHARLES CHENEVIX TRENCH, *A History of Horsemanship* (1970); and GLENN R. VERNAM, *Man on Horseback* (1964, reissued 1972), which includes information on the origin and detail of equipment. (*Horse shows*): R.S. SUMMERHAYS and C.E.G. HOPE, *Horse Shows: The Judges, Stewards, Organizers* (1969); *American Horse Shows Association Rule Book* (biennial); HARLAN C. ABBEY, *Horses and Horse Shows* (1980); JUDY RICHTER, *Horse and Rider: From Basics to Show Competition* (1982); EDWARD HART, *Care and Showing of the Heavy Horse* (1981); and LYNDA BLOOM, *Fitting and Showing the Halter Horse* (1980). (*Rules*): The rules for international competitions are given in publications of the FÉDÉRATION ÉQUESTRE INTERNATIONALE; in BOB PHILLIPS (ed.), *Official Report of the Olympic Games* (1969); and in various publications of the BRITISH HORSE SOCIETY. (*Dressage*): RICHARD L. WÄTJEN, *Dressage Riding: A Guide for the Training of Horse and Rider*, 3rd rev. ed. (1979; originally published in German, 1922; 7th German ed., 1975); HANS HANDLER, *The Spanish Riding School*, trans. from the German (1972); ELWYN HARTLEY EDWARDS, *Saddlery: Modern Equipment for Horse and Stable* (1963, reissued 1973); and *The USCTA Book of Eventing: The Official Handbook of the United States Combined Training Association*, ed. by SALLY O'CONNOR (1982). (*Jumping*): J.A. TALBOT-PONSONBY, *Harmony in Horsemanship* (1964, reissued 1972); F.C. AVIS, *Horses and Show Jumping Dictionary*, 2nd ed., ed. by MARGARET B. SLESSOR (1979); and FEDERICO CAPRILLI, *The Caprilli Papers: Principles of Outdoor Equitation*, trans. and ed. by P. SANTINI (1967).

(A.W.P./C.E.C./Ed.)

Human Rights

It is a common observation, and perhaps even a truism, that human beings everywhere desire to live in a world in which their individual and collective well-being is assured. It also is a common observation that this desire often is painfully frustrated as a result of exploitation, oppression, persecution, and other forms of social deprivation. Deeply rooted in these twin observations are the beginnings of what today are called "human rights" and the national and international legal processes that are associated with them.

This article is divided into the following sections:

Historical development of human rights 656
Origins in ancient Greece and Rome
Natural law transformed into natural rights
"Nonsense upon stilts": the critics of natural rights
The persistence of the notion
Defining human rights 657
The nature of human rights: commonly accepted postulates
The content of human rights: three "generations" of rights
"Intergenerational conflict": legitimacy and priority
International human rights enforcement 659
Treaties, declarations, and agreements before World War II
Human rights in the United Nations
Human rights and the Helsinki process
Regional developments
International human rights in domestic courts
Human rights at the turn of the 21st century 664
Bibliography 664

Historical development of human rights

The expression "human rights" is relatively new, having come into everyday parlance only since the end of World War II. It replaced the phrase "natural rights," which had fallen into disfavour in part because the concept of natural law, to which it was intimately linked, had become a matter of great controversy. It replaced as well the later phrase "the rights of Man," which was not universally understood to include the rights of women.

ORIGINS IN ANCIENT GREECE AND ROME

Most students of human rights trace the origins of the concept to ancient Greece and Rome, where it was closely tied to the natural law doctrines of Stoicism. According to the Stoics, human conduct should be judged according to, and brought into harmony with, the laws of nature. A classic example of this view is given in Sophocles' play *Antigone,* in which the title character, upon being reproached by King Creon for defying his command not to bury her slain brother, asserted that she acted in accordance with the immutable laws of the gods.

Natural law

In part because Stoicism played a key role in its formation and spread, Roman law similarly allowed for the existence of a natural law, and with it certain universal rights that extended beyond the rights of citizenship. According to the Roman jurist Ulpian, for example, natural law was that which nature—not the state—assures to all human beings, Roman citizens or not.

It was not until after the Middle Ages, however, that natural law became associated with natural rights. In ancient and medieval times, doctrines of natural law concerned mainly the duties, rather than the rights, of "Man." Moreover, these doctrines recognized the legitimacy of slavery and serfdom and, in so doing, excluded perhaps the most important ideas of human rights as they are understood today—freedom (or liberty) and equality.

In order for human rights to gain general acceptance as natural rights, therefore, certain basic changes in society were necessary, changes of the sort that took place from the decline of European feudalism through the Renaissance. During this period, resistance to religious intolerance and political and economic bondage, the evident failure of rulers to meet their obligations under natural law, and the unprecedented commitment to individual expression and worldly experience all combined to shift the conception of natural law from *duties* to *rights.* The teachings of St. Thomas Aquinas and Hugo Grotius on the European continent, as well as the Magna Carta (1215), the Petition of Right of 1628, and the English Bill of Rights (1689) were proof of this change. All testified to the increasingly popular view that human beings are endowed with certain eternal and inalienable rights that were not renounced when humankind "contracted" to enter the social from the primitive state and were not diminished by the claim of "the divine right of kings."

NATURAL LAW TRANSFORMED INTO NATURAL RIGHTS

The modern conception of natural law as meaning natural rights was elaborated primarily by thinkers of the 17th and 18th centuries. The many scientific and intellectual achievements of the 17th century encouraged a belief in natural law and universal order, and during the 18th century, the so-called Age of Enlightenment, a growing confidence in human reason and in the perfectibility of human affairs led to the more comprehensive expression of this belief. Particularly important were the writings of John Locke—arguably the most important natural law theorist of modern times—and the works of the 18th-century philosophes centred mainly in Paris, including Montesquieu, Voltaire, and Jean-Jacques Rousseau. Locke argued in detail that certain rights self-evidently pertain to individuals as human beings; that chief among them are the rights to life, liberty (freedom from arbitrary rule), and property; that, upon entering civil society, humankind surrendered to the state—pursuant to a "social contract"—only the right to enforce these natural rights and not the rights themselves; and that the state's failure to secure these rights gives rise to a right to responsible, popular revolution. The philosophes, building on Locke and others, vigorously attacked religious and scientific dogmatism, intolerance, censorship, and social and economic restraints. They sought to discover and act upon universally valid principles governing nature, humanity, and society, including the inalienable "rights of Man," which they treated as a fundamental ethical and social gospel.

Ideas of Locke

Not surprisingly, this liberal intellectual ferment exerted a profound influence in the Western world of the late 18th and early 19th centuries. Together with the Revolution of 1688 in England and the resulting Bill of Rights, it provided the rationale for the wave of revolutionary agitation that swept the West, most notably in North America and France. Thomas Jefferson, who had studied Locke and Montesquieu, gave poetic eloquence to their ideas the Declaration of Independence proclaimed by the 13 American colonies on July 4, 1776: "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the Pursuit of Happiness." Similarly, the Marquis de Lafayette, who won the close friendship of George Washington and who shared the hardships of the American War of Independence, imitated the pronouncements of the English and American revolutions in the Declaration of the Rights of Man and of the Citizen of August 26, 1789, proclaiming that "men are born and remain free and equal in rights" and that "the aim of every political association is the preservation of the natural and imprescriptible rights of man."

In sum, the idea of human rights, though known by another name, played a key role in late 18th- and early 19th-century struggles against political absolutism. It was, indeed, the failure of rulers to respect the principles of freedom and equality that was responsible for this development.

"NONSENSE UPON STILTS": THE CRITICS OF NATURAL RIGHTS

The idea of human rights as natural rights was not without its detractors, however. In the first place, because it was frequently associated with religious orthodoxy, the doctrine of natural rights became less attractive to philosophical and political liberals. Additionally, because they were conceived in essentially absolutist terms, natural rights were increasingly considered to conflict with one another. Most importantly, the doctrine of natural rights came under powerful philosophical and political attack from both the right and the left.

Burke, Hume, and Bentham

In England, for example, conservative political thinkers such as Edmund Burke and David Hume united with liberals such as Jeremy Bentham to condemn the doctrine, the former out of fear that public affirmation of natural rights would lead to social upheaval, the latter out of concern lest declarations and proclamations of natural rights substitute for effective legislation. In his *Reflections on the Revolution in France* (1790), Burke—a believer in natural law who nonetheless denied that the "rights of Man" could be derived from it—criticized the drafters of the Declaration of the Rights of Man and of the Citizen for proclaiming the "monstrous fiction" of human equality, which, he argued, serves but to inspire "false ideas and vain expectations in men destined to travel in the obscure walk of laborious life." Bentham, one of the founders of Utilitarianism, was no less scornful. "Rights," he wrote, "is the child of law; from real law come real rights; but from imaginary laws, from 'law of nature,' come imaginary rights. . . . Natural rights is simple nonsense; natural and imprescriptible rights (an American phrase) . . . [is] rhetorical nonsense, nonsense upon stilts." Agreeing with Bentham, Hume insisted that natural law and natural rights are unreal metaphysical phenomena.

This assault upon natural law and natural rights intensified and broadened during the 19th and early 20th centuries. John Stuart Mill, despite his vigorous defense of liberty, proclaimed that rights ultimately are founded on utility. The German jurist Friedrich Karl von Savigny, England's Sir Henry Maine, and other "historicalist" legal thinkers emphasized that rights are a function of cultural and environmental factors unique to particular communities. The English jurist John Austin argued that the only law is "the command of the sovereign." And the logical positivists of the early 20th century insisted that the pronouncements of ethics were not cognitively significant. By World War I, there were scarcely any theorists who would defend the "rights of Man" in terms of natural law. Indeed, under the influence of 19th-century German Idealism and parallel expressions of rising European nationalism, there were some—the Marxists, for example—who maintained that rights, from whatever source derived, belong preeminently to communities or whole societies and nations.

THE PERSISTENCE OF THE NOTION

Although the heyday of natural rights proved short, the idea of rights nonetheless endured. The abolition of slavery, the implementation of factory legislation, the rise of popular education and trade unionism, the universal suffrage movement—these and other examples of 19th-century reformist impulses afford ample evidence that the idea was not to be extinguished, even if its conceptual foundations had become a matter of general skepticism. But it was not until the rise and fall of Nazi Germany that the idea of human rights truly came into its own. Many of the atrocities committed by the Nazi regime had been officially authorized by Nazi laws and decrees. This fact convinced many that law and morality cannot be grounded in any purely Idealist, Utilitarian, or other consequentialist doctrine. Certain actions, according to this view, are absolutely wrong, no matter what the circumstances.

Human rights in the 20th century

Today, the vast majority of legal scholars and philosophers—particularly in the West—agree that every human being has, at least in theory, some basic rights. Indeed, except for some essentially isolated late 19th-century and early 20th-century demonstrations of international humanitarian concern to be noted below, the last half of the 20th century may fairly be said to mark the birth of the international as well as the universal recognition of human rights. In the treaty charter establishing the United Nations, for example, all member states pledged themselves to take joint and separate action for the achievement of "universal respect for, and observance of, human rights and fundamental freedoms for all without distinction as to race, sex, language, or religion." Similarly, in the Universal Declaration of Human Rights (1948), representatives from many cultures endorsed the rights therein set forth "as a common standard of achievement for all peoples and all nations." And in 1976, the International Covenant on Economic, Social and Cultural Rights and the International Covenant on Civil and Political Rights entered into force and effect.

Defining human rights

To say that there is widespread acceptance of the principle of human rights is not to say that there is complete agreement about the nature and scope of such rights. Among the basic questions that have yet to receive conclusive answers are the following: Whether human rights are to be viewed as divine, moral, or legal entitlements; whether they are to be validated by intuition, culture, custom, social contract, principles of distributive justice, or as prerequisites for happiness; whether they are to be understood as irrevocable or partially revocable; and whether they are to be broad or limited in number and content.

THE NATURE OF HUMAN RIGHTS: COMMONLY ACCEPTED POSTULATES

Despite this lack of consensus, a number of widely accepted—and interrelated—postulates can assist in the task of defining human rights. Five in particular stand out, though not even these are without controversy.

First, regardless of their ultimate origin or justification, human rights are understood to represent both individual and group demands for political power, wealth, education, and other social goods and benefits, the most fundamental of which is respect and its constituent elements of reciprocal tolerance and mutual forbearance in the pursuit of all other goods. Consequently, human rights imply claims against persons and institutions impeding the achievement of these goods, as well as standards for judging the legitimacy of laws and traditions. At bottom, human rights qualify state sovereignty and power, sometimes expanding the latter even while circumscribing the former.

Second, human rights are commonly assumed to refer, in some vague sense, to "fundamental" as distinct from "nonessential" goods or benefits. In fact, some theorists go so far as to limit human rights to a single core right or two—for example, the right to life or the right to equal freedom of opportunity. The tendency is to emphasize "basic needs" and to rule out "mere wants."

Third, reflecting varying environmental circumstances, differing worldviews, and inescapable interdependencies between goods and benefits, human rights refer to a wide continuum of claims, ranging from the most justiciable to the most aspirational. Human rights partake of both the legal and the moral orders, sometimes indistinguishably. They are expressive of both the "is" and the "ought" in human affairs.

Fourth, most assertions of human rights—though arguably not all—are qualified by the limitation that the rights of any particular individual or group in any particular instance are restricted as much as is necessary to secure the comparable rights of others and the aggregate common interest. Given this limitation, human rights are sometimes designated prima facie rights, so that ordinarily it makes little or no sense to think of them in absolutist terms.

Universal nature of human rights

Finally, human rights are understood to be quintessentially universal in character, in some sense equally pos-

sessed by all human beings everywhere, including in certain instances even the unborn. In stark contrast to the divine right of kings and other such conceptions of privilege, human rights extend in theory to every person on Earth without regard to merit or need, simply for being human.

In several critical respects, however, all of these postulates raise more questions than they answer. Granted that human rights qualify state power, do they also qualify private power? If so, when and how? What does it mean to say that a right is fundamental? What is the value of embracing non-justiciable rights as part of the jurisprudence of human rights? When and according to what criteria does the right of one person or group of people give way to the right of another? What happens when individual and group rights collide? How are universal human rights determined? Are they a function of culture or ideology, or are they determined according to some transnational consensus? If the latter, a regional consensus? A global consensus? And if a regional or global consensus, how exactly would it be ascertained, and how would it be reconciled with the right of nations and peoples to self-determination? Is the existence of universal human rights incompatible with the notion of national sovereignty? Should supranational institutions have the power to nullify local, regional, and national laws on capital punishment, "honor killing," veil wearing, female genital cutting, and other practices? How would such a situation comport with Western conceptions of democracy and representative government? In other words, however accurate, the five foregoing postulates are fraught with questions about the content and legitimate scope of human rights and about the priorities, if any, that exist among them. Like the issue of the origin and justification of human rights, all five are controversial.

THE CONTENT OF HUMAN RIGHTS: THREE "GENERATIONS" OF RIGHTS

Like all normative traditions, the human rights tradition is a product of its time. Therefore, to understand better the debate over the content and scope of human rights and the priorities claimed among them, it is useful to note the dominant schools of thought that have informed the human rights tradition since the beginning of modern times.

Particularly helpful in this regard is the notion of "three generations of human rights" advanced by the French jurist Karel Vasak. Inspired by the three normative themes of the French Revolution, they are: the first generation of civil and political rights (*liberté*); the second generation of economic, social, and cultural rights (*égalité*); and the third generation of solidarity rights (*fraternité*). Vasak's model is, of course, a simplified expression of an extremely complex historical record, and it is not intended to suggest a linear process in which each generation gives birth to the next and then dies away. The three generations are cumulative, overlapping, and, it is important to note, interdependent.

Liberté: **civil and political rights.** The first generation of civil and political rights derives primarily from the 17th- and 18th-century reformist theories noted above—*i.e.,* those associated with the English, American, and French revolutions. Infused with the political philosophy of liberal individualism and the related economic and social doctrine of laissez-faire, the first generation conceives of human rights more in negative ("freedoms from") than positive ("rights to") terms; it favours the abstention over the intervention of government in the quest for human dignity. Belonging to this first generation, thus, are rights such as those set forth in Articles 2–21 of the Universal Declaration of Human Rights, including freedom from gender, racial, and equivalent forms of discrimination; the right to life, liberty, and the security of the person; freedom from slavery or involuntary servitude; freedom from torture and from cruel, inhuman, or degrading treatment or punishment; freedom from arbitrary arrest, detention, or exile; the right to a fair and public trial; freedom from interference in privacy and correspondence; freedom of movement and residence; the right to asylum from persecution; freedom of thought, conscience, and religion; freedom of opinion and expression; freedom of peaceful assembly and association; and the right to participate in government, directly or through free elections. Also included are the right to own property and the right not to be deprived of it arbitrarily, which were fundamental to the interests fought for in the American and French revolutions and to the rise of capitalism.

"Negative" rights

It should be noted that these and other first-generation rights do not correspond completely to the idea of "negative" rights. The right to security of the person, to a fair and public trial, to asylum from persecution, and to free elections, for example, manifestly cannot be assured without some affirmative government action. What is constant in this first-generation conception is the notion of liberty, a shield that safeguards the individual—alone and in association with others—against the abuse of political authority. This is the core value. Featured in the constitution of almost every country in the world and dominating the majority of international declarations and covenants adopted since World War II, this essentially Western liberal conception of human rights is sometimes romanticized as a triumph of Hobbesian–Lockean individualism over Hegelian statism.

Égalité: **economic, social, and cultural rights.** The second generation of economic, social, and cultural rights finds its origins primarily in the socialist tradition that was foreshadowed among the Saint-Simonians of early 19th-century France and variously promoted by revolutionary struggles and welfare movements that have taken place since. In large part, it is a response to the abuses of capitalist development and its underlying, essentially uncritical, conception of individual liberty that tolerated, even legitimated, the exploitation of working classes and colonial peoples. Historically, it is a counterpoint to the first generation of civil and political rights, with human rights conceived more in positive than in negative terms and requiring more the intervention than the abstention of the state for the purpose of assuring equitable distribution of the goods and benefits involved. Illustrative are some of the rights set forth in Articles 22–27 of the Universal Declaration of Human Rights, such as the right to social security; the right to work and to protection against unemployment; the right to rest and leisure, including periodic holidays with pay; the right to a standard of living adequate for the health and well-being of self and family; the right to education; and the right to the protection of one's scientific, literary, and artistic production.

"Positive" rights

But in the same way that all the rights embraced by the first generation of civil and political rights cannot properly be designated "negative rights," so all the rights embraced by the second generation of economic, social, and cultural rights cannot properly be labeled "positive rights." For example, the right to free choice of employment, the right to form and to join trade unions, and the right freely to participate in the cultural life of the community (found in Articles 23 and 27) do not inherently require affirmative state action to ensure their enjoyment. Nevertheless, most of the second-generation rights do necessitate state intervention because they subsume demands more for material than for intangible goods and benefits. Second-generation rights are, fundamentally, claims to social equality. Partly because of the comparatively late arrival of socialist-communist and compatible "Third World" influence in the normative domain of international affairs, however, the internationalization of these rights has been relatively slow in coming; and with free-market capitalism in ascendancy under the banner of "globalization" at the turn of the 21st century, it is not likely that these rights will come of age any time soon. On the other hand, as the social inequities created by unregulated capitalism become more and more evident over time and are not accounted for by sex or race discrimination, it is probable that the struggle for second-generation rights will grow and mature. This tendency is already apparent in the evolving European Union.

Fraternité: **solidarity rights.** Finally, the third generation of solidarity rights is best understood as a product of both the rise and the decline of the nation-state in the last half of the 20th century. Of the six rights in this group, three reflect the emergence of "Third World" nationalism and its "revolution of rising expectation"—*i.e.,* its demand for a global redistribution of power, wealth, and other important goods and benefits: the right to political, economic, social, and cultural self-determination; the right to economic and

social development; and the right to participate in and benefit from "the common heritage of mankind" (shared Earth–space resources; scientific, technical, and other information and progress; and cultural traditions, sites, and monuments). The other three third- generation rights—the right to peace, the right to a healthy and sustainable environment, and the right to humanitarian disaster relief—suggest the impotence or inefficiency of the nation-state in certain critical respects.

Collective rights

All six of these rights tend to be posed as collective rights, requiring the concerted efforts of all social forces, to a substantial degree on a planetary scale. Nevertheless, each of them also manifests an individual dimension. For example, while it may be said to be the collective right of all countries and peoples (especially developing countries and non-self-governing peoples) to secure a "new international economic order" that would eliminate obstacles to their economic and social development, so also may it be said to be the individual right of every person to benefit from a developmental policy that is based on the satisfaction of material and nonmaterial human needs. It is also important to note that the majority of these rights are more aspirational than justiciable in character, and their status as international human rights norms remains ambiguous.

"INTERGENERATIONAL CONFLICT": LEGITIMACY AND PRIORITY

***Liberté* versus *égalité*.** The fact that human rights have been defined so broadly should not be taken to imply that the three generations of rights are equally acceptable to everyone. Nor should it suggest that they or their separate elements have been greeted with equal urgency. The debate about the nature and content of human rights reflects, after all, a struggle for power and for favoured conceptions of the "good society."

First-generation proponents, for example, are inclined to exclude second- and third-generation rights from their definition of human rights altogether (or, at best, to label them as "derivative"). In part this is because of the complexities involved in putting these rights into operation. First-generation rights are viewed as more feasible because they stress the absence over the presence of government, and feasibility is transformed into a prerequisite of a comprehensive definition of human rights, so that aspirational claims to entitlement are deemed not to be rights at all. The most forceful explanation, however, is more ideologically or politically motivated. Persuaded that egalitarian claims against the rich, particularly where collectively espoused, are unworkable without a severe decline in liberty and equality, first-generation proponents, inspired by the natural law and laissez-faire traditions, are partial to the view that human rights are inherently independent of organized society and are individualistic.

Conversely, second- and third-generation defenders often look upon first-generation rights, at least as commonly practiced, as insufficiently attentive to material—especially "basic"—human needs and, indeed, as instruments in the service of unjust social orders, hence constituting a "bourgeois illusion." Accordingly, if they do not place first-generation rights outside their definition of human rights, they tend to assign such rights a low status and to treat them as long-term goals that will come to pass only after the imperatives of economic and social development have been met, to be realized gradually and fully achieved only sometime vaguely in the future.

This liberty-equality and individualist-collectivist debate, it must be added, was especially evident during the Cold War, reflecting the tensions that then existed between Liberal and Marxist conceptions of sovereign public order. Different conceptions of rights contain the potential for challenging the legitimacy and supremacy not only of one another but, more importantly, of the sociopolitical systems with which they are most intimately associated.

The end of the Cold War

The relevance of custom and tradition. With the end of the Cold War, however, the debate took on a more North-South character and was supplemented by a cultural-relativist critique that eschewed the universality of human rights doctrines, principles, and rules on the grounds that they were Western in origin and therefore of limited relevance in non-Western settings. The viewpoint underlying this assertion—that the scope of human rights in any given society is fundamentally determined by local, national, or regional customs and traditions—may seem problematic, especially when one considers that the idea of human rights and many of its precepts are found in all the great philosophical and religious traditions. Nevertheless, the historical development of human rights demonstrates that it cannot be wholly mistaken. Nor is it surprising that it should emerge soon after the end of the Cold War. First prominently expressed at an Asian preparatory meeting to the second UN World Conference on Human Rights convened in Vienna in June 1993, it reflected the end of a bipolar system of alliances that had discouraged independent foreign policies and minimized cultural and political differences among countries allied to the same superpower. Against the backdrop of increasing human rights interventionism on the part of the UN, regional organizations, and deputized coalitions of states (as in Bosnia-Herzegovina, Somalia, Liberia, Rwanda, and Haiti, for example), the viewpoint served as a functional equivalent of the doctrine of respect for national sovereignty and territorial integrity, a doctrine whose influence had been declining not only in human rights affairs but also in affairs related to national security, economics, and the environment. As a consequence, there remains sharp disagreement about the legitimate scope of human rights and about the priorities that are claimed among them.

Inherent risks of the debate. On final analysis, however, this legitimacy-priority debate can be misleading. Although useful for pointing out how notions of liberty and individualism have been used to rationalize the abuses of capitalism and how notions of equality, collectivism, and culture have been alibis for authoritarian governance, in the end it risks obscuring at least three essential truths that must be taken into account if the contemporary worldwide human rights movement is to be objectively understood.

First, one-sided characterizations of legitimacy and priority are very likely, at least over the long term, to undermine the credibility of their proponents and the defensibility of the rights they regard as preeminently important. Second, such characterizations do not accurately reflect reality. In the real world, essentially individualistic societies tolerate, and even promote, certain collectivist values, and essentially communal societies tolerate, and even promote, certain individualistic values.

Finally, none of the international human rights instruments currently in force or proposed says anything about the legitimacy or priority of the rights they address, save possibly in the case of rights that by international covenant are stipulated to be "nonderogable" and therefore, arguably, more fundamental than others (*e.g.*, freedom from arbitrary or unlawful deprivation of life). To be sure, some disagreements about legitimacy and priority can derive from differences of definition (*e.g.*, what is "torture" or "inhuman treatment" to one may not be to another). Similarly, disagreements also can arise when treating the problem of implementation. For example, some insist first on certain civil and political guarantees, whereas others defer initially to conditions of material well-being. Such disagreements, however, reflect differences in political agendas and have little if any conceptual utility. As indicated by numerous resolutions of the UN General Assembly and the Declaration and Programme of Action of the 1993 Vienna World Conference on Human Rights, there is a growing consensus that human rights form an indivisible whole and that the protection of human rights should not be a matter of purely national jurisdiction. The extent to which the international community actually protects the rights it prescribes, on the other hand, is a different matter.

International human rights enforcement

TREATIES, DECLARATIONS, AND AGREEMENTS BEFORE WORLD WAR II

Ever since ancient times, but especially since the emergence of the modern state system, the Age of Discovery, and the accompanying spread of industrialization and European culture throughout the world, there has developed,

for economic and other reasons, a unique set of customs and conventions regarding the humane treatment of foreigners. This evolving International Law of State Responsibility for Injuries to Aliens, as these customs and conventions came to be called, represents the beginning of active concern for human rights on the international plane. The founding fathers of international law—particularly Francisco de Vitoria (1486?–1546), Hugo Grotius (1583–1645), and Emmerich de Vattel (1714–67)—were quick to observe that all persons, outlander as well as other, were entitled to certain natural rights; and they emphasized, consequently, the importance of according aliens fair treatment.

With the exception of occasional treaties to secure the protection of Christian denominations, it was not until the start of the 19th century that active international concern for the rights of nationals began to make itself felt. Then, in the century and a half before World War II, several noteworthy efforts to encourage respect for nationals by international means began to shape what today is called the International Law of Human Rights.

International Law of Human Rights

Throughout the 19th and early 20th centuries, numerous military operations and diplomatic representations, not all of them with the purest of motives but performed nonetheless in the name of "humanitarian intervention" (a customary international law doctrine), undertook to protect oppressed and persecuted minorities in the Ottoman Empire and in Syria, Crete, various Balkan countries, Romania, and Russia. Paralleling these actions, first at the Congress of Vienna (1814–15) and later between the two world wars, a series of treaties and international declarations sought the protection of certain racial, religious, and linguistic minorities in central and eastern Europe and in the Middle East. During the same period, the movement to combat and suppress slavery and the slave trade found expression in treaties involving the major commercial powers, beginning with the Treaty of Paris (1814) and culminating in the International Slavery Convention (1926).

In addition, toward the end of the 19th century and continuing well beyond World War II, the community of nations, inspired largely by persons associated with what is now the International Committee of the Red Cross, concluded a series of multilateral declarations and agreements designed to temper the conduct of hostilities, protect the victims of war, and otherwise elaborate the humanitarian law of war (now commonly referred to as "international humanitarian law"). At about the same time, first with two multilateral labour conventions concluded in 1906 and subsequently at the initiative of the International Labour Organisation (ILO; established in 1919), a reformist-minded international community embarked upon a variety of collaborative measures directed at the promotion of human rights. These included not only fields traditionally associated with labour law and labour relations but also—mainly after World War II—such core human rights concerns as forced labour, discrimination in employment and occupation, freedom of association for collective bargaining, and equal pay for equal work.

Finally, during the interwar period, the Covenant establishing the League of Nations (1919)—though not formally recognizing "the rights of Man" and failing to lay down a principle of racial nondiscrimination as requested by Japan—nevertheless committed the League's members to several human rights goals: fair and humane working conditions; the execution of agreements regarding traffic in women and children; the prevention and control of disease in matters of international concern; and the just treatment of indigenous colonial peoples. Also, the victorious powers—who as "mandatories" were entrusted by the League with the tutelage of colonies formerly governed by Germany and Turkey—accepted responsibility for the well-being and development of the inhabitants of those territories as "a sacred trust of civilization." This arrangement was later carried over into the trusteeship system of the United Nations.

Human rights goals of the League of Nations

As important as these efforts were, however, it was not until after the war—and the Nazi atrocities accompanying it—that active concern for human rights truly came of age internationally. In the proceedings of the International Military Tribunal at Nürnberg in 1945–46, German high officials were tried not only for "crimes against peace" and "war crimes" but also for "crimes against humanity" committed against civilian populations, even if the crimes were in accordance with the laws of the country where perpetrated. Although the tribunal, whose establishment and rulings subsequently were endorsed by the UN General Assembly, applied a cautious approach to allegations of crimes against humanity, it nonetheless made the treatment by a state of its own citizens the subject of international criminal process. The ad hoc international criminal tribunals established in 1993–94 for the prosecution of serious violations of international humanitarian law in the former Yugoslavia and Rwanda, as well as the permanent International Criminal Court created in 1998, are its first heirs on the international plane. In June 2001 the International Criminal Tribunal for Yugoslavia took custody of former Yugoslav president Slobodan Milošević and charged him with war crimes and crimes against humanity for actions allegedly committed by Serbian forces in Kosovo in 1999.

HUMAN RIGHTS IN THE UNITED NATIONS

The Charter of the United Nations (1945) begins by reaffirming a "faith in fundamental human rights, in the dignity and worth of the human person, in the equal rights of men and women and of nations large and small." It states that the purposes of the UN are, among other things, "to develop friendly relations among nations based on respect for the principle of equal rights and self-determination of peoples . . . [and] to achieve international co-operation . . . in promoting and encouraging respect for human rights and for fundamental freedoms for all without distinction as to race, sex, language, or religion." In addition, in two key articles all members "pledge themselves to take joint and separate action in cooperation with the Organization" for the achievement of these and related purposes. It should be noted, however, that a proposal to ensure the protection as well as the promotion of human rights was explicitly rejected at the Charter-drafting San Francisco conference establishing the UN. Moreover, the Charter expressly provides that nothing in it "shall authorize the United Nations to intervene in matters which are essentially within the domestic jurisdiction of any state," except upon a Security Council finding of a "threat to the peace, breach of the peace, or act of aggression." Although typical of major constitutive instruments, the Charter is conspicuously given to generality and vagueness in its human rights clauses, among others.

Protection versus promotion of human rights

Thus, not surprisingly, the reconciliation of the Charter's human rights provisions with the history of its drafting and its "domestic jurisdiction" clause has given rise to legal and political controversy. Some authorities have argued that, in becoming parties to the Charter, states accept no more than a nebulous promotional obligation toward human rights and that, in any event, the UN has no standing to insist on human rights safeguards in member states. Others have insisted that the Charter's human rights provisions, being part of a legally binding treaty, clearly involve some element of legal obligation; that the "pledge" made by states upon becoming party to the Charter consequently represents more than a moral statement; and that the "domestic jurisdiction" clause does not apply because human rights no longer can be considered matters "essentially within the domestic jurisdiction" of states.

When all is said and done, however, it is clear from the actual practice of the UN that the problem of resolving these opposing contentions has proved less formidable than the statements of governments and the opinions of scholars would suggest. Neither the Charter's drafting history nor its "domestic jurisdiction" clause—nor, indeed, its generality and vagueness in respect of human rights—has prevented the UN from evaluating and responding to specific human rights situations—*e.g.*, as in the Security Council's imposition of a mandatory arms embargo against South Africa in 1977 and its authorization of the use of military force to end human rights abuses in Somalia and Haiti in the early 1990s. Of course, governments usually are protective of their sovereignty, or domestic ju-

Intervention in Somalia and Haiti

risdiction. Also, the UN organs responsible for the promotion and protection of human rights suffer from most of the same disabilities that afflict the UN as a whole, in particular the absence of supranational authority and the presence of divisive power politics. Hence, it cannot be expected that UN actions in defense of human rights will be, normally, either swift or categorically effective. Indeed, on the basis of the historical record to date, it may fairly be said that serious UN efforts at human rights implementation are often thwarted, not least at the hands of the major powers. In 1999, for example, opposition by Russian and China prevented the Security Council from agreeing on forceful measures to end the persecution by Serbia of ethnic Albanians in the province of Kosovo, prompting the United States and other members of the North Atlantic Treaty Organization (NATO) to take matters into their own hands through a massive bombing campaign against Serbian targets. Nevertheless, assuming some political will, the legal obstacles to UN enforcement of human rights are not insurmountable.

Primary responsibility for the promotion and protection of human rights under the UN Charter rests in the General Assembly and, under its authority, in the Economic and Social Council (ECOSOC), the Commission on Human Rights, and the UN High Commissioner for Human Rights (UNHCHR). The UN Commission on Human Rights, an intergovernmental subsidiary body of ECOSOC that met for the first time in 1947, serves as the UN's central policy organ in the human rights field. The UNHCHR, a post created by the General Assembly in 1993, is the official principally responsible for implementing and coordinating UN human rights programs and projects, including overall supervision of the UN's Geneva-based Centre for Human Rights, a bureau of the UN Secretariat.

UN Commission on Human Rights and its instruments. For the first 20 years of its existence (1947–1966), the UN Commission on Human Rights concentrated its efforts on setting human rights standards. Together with other UN bodies such as the ILO, the UN Educational, Scientific and Cultural Organization (UNESCO), the UN Commission on the Status of Women, and the Commission on Crime Prevention and Criminal Justice, it has drafted standards and prepared a number of international human rights instruments. Among the most important of these have been the Universal Declaration of Human Rights (1948), the International Covenant on Economic, Social and Cultural Rights (1966), and the International Covenant on Civil and Political Rights together with its Optional Protocols (1966; 1989). Collectively known as the "International Bill of Human Rights," these three instruments serve as touchstones for interpreting the human rights provisions of the UN charter. Also central have been the International Convention on the Elimination of All Forms of Racial Discrimination (CERD; 1965), the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW; 1979), the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (1984), and the Convention on the Rights of the Child (1989), each elaborating on provisions of the International Bill of Human Rights.

The "International Bill of Human Rights"

The Commission continues to perform this standard-setting role. Beginning in 1967, however, it was specifically authorized to deal with violations of human rights, and since then it has set up elaborate mechanisms and procedures to investigate alleged violations of human rights and otherwise monitor compliance by states with international human rights law. Thus, much of the work of the Commission is now investigatory, evaluative, and advisory in character. It annually establishes a working group to consider and make recommendations concerning alleged "gross violations" of human rights referred to it by its Sub-Commission on Prevention of Discrimination and Protection of Minorities. Also, on an ad hoc basis, the commission appoints Special Rapporteurs, Special Representatives, Special Committees, and other envoys to examine human rights situations and report back to the Commission. These fact-finding and implementation mechanisms and procedures were the focus of the Commission's attention during the 1970s and '80s. In the 1990s the Commission increasingly turned to economic, social, and cultural rights, including the right to development and the right to an adequate standard of living. Particular attention has been paid to the rights of minorities, indigenous peoples, women, and children.

The UN High Commissioner for Human Rights. Appointed by the secretary-general in a regular rotation of geographic regions and approved by the General Assembly, the UN High Commissioner for Human Rights (UNHCHR) serves a fixed term of four years with the possibility of renewal for an additional four-year term. The first high commissioner, José Ayala Lasso of Ecuador, took office in April 1994 and was temporarily succeeded by Ralph Zacklin of the United Kingdom in 1997–98. Mary Robinson, former president of Ireland, was the second high commissioner. Among other duties, the high commissioner is charged by the General Assembly with promoting and protecting all civil, political, economic, social, and cultural rights; providing advisory services and technical and financial assistance in the field of human rights to states that request them; coordinating human rights promotion and protection activities throughout the UN system, including education and public information programs; and enhancing international cooperation for the promotion and protection of human rights.

The Universal Declaration of Human Rights. The catalog of rights set out in the Universal Declaration of Human Rights, which was adopted without dissent by the General Assembly on Dec. 10, 1948, is scarcely less than the sum of most of the important traditional political and civil rights of national constitutions and legal systems, including equality before the law; protection against arbitrary arrest; the right to a fair trial; freedom from ex post facto criminal laws; the right to own property; freedom of thought, conscience, and religion; freedom of opinion and expression; and freedom of peaceful assembly and association. Also enumerated are such economic, social, and cultural rights as the right to work, the right to form and join trade unions, the right to rest and leisure, the right to a standard of living adequate for health and well-being, and the right to education.

The Universal Declaration, it should be noted, is not a treaty. It was meant to proclaim "a common standard of achievement for all peoples and all nations" rather than enforceable legal obligations. Nevertheless, the Universal Declaration has acquired a status juridically more important than originally intended, and it has been widely used, even by national courts, as a means of judging compliance with human rights obligations under the UN Charter.

Juridical status of the Universal Declaration

The International Covenant on Civil and Political Rights and its Optional Protocols. The civil and political rights guaranteed by the International Covenant on Civil and Political Rights, which was opened for signature on Dec. 19, 1966, and entered into force on March 23, 1976, incorporate almost all those proclaimed in the Universal Declaration, including the right to nondiscrimination, though not the right to own property and the right to asylum. The Covenant also designates several rights that are not listed in the Universal Declaration, among them the right of all peoples to self-determination and the right of ethnic, religious, or linguistic minorities to enjoy their own culture, to profess and practice their own religion, and to use their own language. To the extent that the Universal Declaration and the Covenant overlap, however, the latter is understood to explicate and help interpret the former.

In addition, the Covenant calls for the establishment of a Human Rights Committee to study reports submitted by the state parties on measures they have adopted to give effect to the rights recognized in the Covenant. For state parties that have expressly recognized the competence of the Committee in this regard, the Committee also may respond to allegations by one state party that another state party is not fulfilling its obligations under the Covenant. If the Committee is unable to resolve the problem, the matter is referred to an ad hoc conciliation commission, which eventually reports its findings on all questions of fact, plus its views on the possibilities of an amicable solution. State parties that become party to the Covenant's first Optional Protocol further recognize the competence of the Human

Rights Committee to consider and act upon communications from individuals claiming to be victims of Covenant violations. Other treaty-based organs within the UN system that are empowered to consider grievances from individuals are the Committee on the Elimination of Racial Discrimination and the Committee on Torture.

Abolition of the death penalty

The Covenant's Second Optional Protocol is aimed at abolishing the death penalty worldwide. Adopted in 1989 and entered into force in 1991, it has been favourably received in most of the countries of western Europe and many countries in Latin America.

The International Covenant on Economic, Social and Cultural Rights. Just as the International Covenant on Civil and Political Rights elaborates upon most of the civil and political rights enumerated in the Universal Declaration of Human Rights, so the International Covenant on Economic, Social and Cultural Rights elaborates upon most of the economic, social, and cultural rights set forth in the Universal Declaration: the right to work; the right to just and favourable conditions of work; trade union rights; the right to social security; rights relating to the protection of the family; the right to an adequate standard of living; the right to health; the right to education; and rights relating to culture and science. Unlike its companion International Covenant on Civil and Political Rights, however, this covenant generally is not geared toward immediate implementation, the state parties having agreed only "to take steps" toward "achieving progressively the full realization of the rights recognized in the . . . Covenant," and then subject to "the maximum of [their] available resources." One obligation, however, is subject to immediate application: the prohibition of discrimination in the enjoyment of the rights enumerated on grounds of race, colour, sex, language, religion, political or other opinion, national or social origin, property, and birth or other status. Also, the international supervisory measures that apply to the Covenant oblige the state parties to report to the UN Economic and Social Council on the steps they have adopted and the progress they have made in achieving the realization of the enumerated rights.

Other UN human rights conventions and declarations. Numerous other human rights treaties drafted under UN auspices address a broad range of concerns, including the prevention and punishment of the crime of genocide; the humane treatment of military and civilian personnel in time of war; the status of refugees; the protection and reduction of stateless persons; the abolition of slavery, forced labour, and discrimination in employment and occupation; the suppression and punishment of the crime of apartheid; the elimination of discrimination in education; the promotion of the political rights of women; the protection of minorities and indigenous peoples; and the promotion of equality of opportunity and treatment among migrant workers. In addition to overseeing human rights treaties, the UN also adopts declarations, in the form of resolutions, aimed at promoting human rights. Although technically not binding on member states in the sense of a treaty, such declarations may nevertheless create strong expectations about authority and control. Perhaps the best-known examples subsequent to the Universal Declaration are the Declaration on the Granting of Independence to Colonial Countries and Peoples (1960) and the Declaration on Principles of International Law concerning Friendly Relations and Co-operation among States in accordance with the Charter of the United Nations (1970). Other declarations have addressed the rights of disabled persons; the elimination of all forms of intolerance and of discrimination based on religion or belief; the right of peoples to peace; the right to development; the rights of persons belonging to national or ethnic, religious, and linguistic minorities; and the elimination of violence against women.

HUMAN RIGHTS AND THE HELSINKI PROCESS

After World War II, international concern for human rights was evident at the global level outside the UN as well as within it, most notably in the proceedings and aftermath of the Conference on Security and Co-operation in Europe (CSCE), convened in Helsinki, Finland, on July 3, 1973, and concluded there (after continuing deliberations in Geneva) on August 1, 1975. Attended by representatives of 35 governments—including the NATO countries, the Warsaw Pact nations, and 13 neutral and nonaligned European states—the Conference had as its principal purpose a mutually satisfactory definition of peace and stability between East and West, previously made impossible by the Cold War. In particular, the former Soviet Union wished to gain recognition of its western frontiers as established at the end of World War II (which ended without the conclusion of an omnibus peace treaty). The West, with no realistic territorial claims of its own, sought concessions primarily on security requirements and human rights.

The Helsinki Final Act

The Final Act of the conference, the importance of which is reflected in its having been signed by almost all of the principal governmental leaders of the day, begins with a Declaration on Principles Guiding Relations between Participating States, in which the "Participating States" solemnly declare "their determination to respect and put into practice," alongside other "guiding" principles, "respect [for] human rights and fundamental freedoms, including the freedom of thought, conscience, religion or belief" and "respect [for] the equal rights of peoples and their right to self-determination." It was hoped that this declaration would mark the beginning of a liberalization of authoritarian regimes.

From the earliest discussions, however, it was clear that the Helsinki Final Act was not intended as a legally binding instrument. "Determination to respect" and "put into practice" were treated as moral commitments only; the Declaration of Principles was said not to prescribe international law; and nowhere did the participants provide for enforcement machinery. On the other hand, the Declaration of Principles, including its human rights principles, was always viewed as being at least consistent with international law; and in providing for periodic follow-up conferences, it made possible a unique negotiating process ("the Helsinki process") to review compliance with its terms, thus creating expectations concerning the conduct of the "Participating States." In these ways it proved to be an important force in the collapse of the Iron Curtain and the transformation of eastern Europe.

The Helsinki process served also to establish a mechanism for the evolution of the CSCE in the 1990s from a forum for discussion to an operational institution. In 1994 the CSCE was renamed the Organization for Security and Co-operation in Europe (OSCE), and its principal organs and bureaus now include an Office for Democratic Institutions and Human Rights, a Conflict Prevention Centre, a High Commissioner on National Minorities, and a Court of Conciliation and Arbitration. These offices have been increasingly pressed into service to alleviate major deprivations of human rights, particularly those arising from ethnic conflicts. In addition, the Vienna Human Dimension Mechanism and the Moscow Human Dimension Mechanism provide a preliminary formal means of raising resolving disputes about violations of human rights commitments, including the possibility of on-site investigation by independent experts. It should be noted, however, that all these mechanisms bespeak an essentially interstate process; neither individuals nor NGOs have access to them except indirectly as suppliers of information and conveyors of political pressure. They thus contrast markedly with the individual complaint procedures that are available within the UN and in regional human rights systems.

REGIONAL DEVELOPMENTS

Action for the international promotion and protection of human rights has proceeded at the regional level in Europe, the Americas, Africa, the Middle East, and, to a minor extent, Asia. Only the first three of these regions, however, have created enforcement mechanisms within the framework of a human rights charter.

European human rights system. On Nov. 4, 1950, the Council of Europe agreed to the European Convention for the Protection of Human Rights and Fundamental Freedoms, the substantive provisions of which are based on a draft of what is now the International Covenant on Civil and Political Rights. Together with its 11 additional protocols, this convention, which entered into force on Sep-

tember 3, 1953, represents the most advanced and successful international experiment in the field to date. Over the years, the enforcement mechanisms created by the Convention have developed a considerable body of case law on questions regulated by the Convention, which the state parties typically have honoured and respected. In some European states the provisions of the Convention are deemed to be part of domestic constitutional or statutory law. Where this is not the case, the state parties have taken other measures to make their domestic laws conform with their obligations under the Convention.

Notwithstanding these successes, a significant streamlining of the European human rights regime took place on Nov. 1, 1998, when Protocol No. 11 to the Convention entered into force. Pursuant to the protocol, the enforcement mechanisms created by the Convention, the European Commission of Human Rights and the European Court of Human Rights, were merged into a reconstituted court, which was now empowered to hear individual (as opposed to interstate) petitions or complaints without the prior approval of the local government. The court's decisions are final and binding on the state parties to the Convention.

European Social Charter

A companion instrument to the European Convention is the European Social Charter (1961) and its additional protocol (1988). In contrast to the adjudicatory enforcement procedures of the European Convention, the charter's provisions are implemented through an elaborate system of control based on progress reports to the various committees and organs of the Council of Europe.

Inter-American human rights system. In 1948, concurrent with its establishment of the Organization of American States (OAS), the Ninth Pan-American Conference adopted the American Declaration on the Rights and Duties of Man, which, unlike the Universal Declaration of the UN, set out the duties as well as the rights of individual citizens. In 1959, a meeting of the American Ministers for Foreign Affairs created the Inter-American Commission on Human Rights, which has since undertaken important investigative activities. Finally, in 1969, the Inter-American Specialized Conference on Human Rights adopted the American Convention on Human Rights, which, among other things, established the Inter-American Court of Human Rights, which sits in San José, Costa Rica. Of the 26 western hemispheric states that so far have signed the convention, only the United States has yet to ratify it.

The core structure of the Inter-American human rights system is similar to that of its European counterpart. Nevertheless, some noteworthy differences exist, and three stand out in particular. First, reminiscent of the American Declaration on the Rights and Duties of Man, the American convention details individual duties as well as individual rights. Second, the American convention, unlike the European convention prior to Protocol No. 11, gives primary consideration to individual rather than interstate petitions or complaints. Finally, both the Inter-American Commission and the Inter-American Court operate beyond the framework of the American convention. The commission is as much an organ of the OAS Charter as of the American convention, and its powers and procedures can vary significantly depending on which body authorizes its activities. The court, though primarily an organ of the convention, has jurisdiction to interpret human rights provisions of other treaties, including the OAS Charter.

The African Charter

African human rights system. In 1981, the Eighteenth Assembly of Heads of State and Government of the Organization of African Unity (OAU) adopted the African Charter on Human and Peoples' Rights. Also known as the "Banjul Charter" for having been drafted in Banjul, Gambia, it entered into force on October 21, 1986, and boasts the vast majority of the states of Africa as parties.

Like its American and early European counterparts, the African Charter provides for a human rights commission, which has both promotional and protective functions. There is no restriction on who may file a complaint with it. In contrast to the European and American procedures, however, concerned states are encouraged to reach a friendly settlement without formally involving the investigative or conciliatory mechanisms of the Commission. Also, the African Charter does not call for a human rights court. African customs and traditions, it has been said, emphasize mediation, conciliation, and consensus rather than the adversarial and adjudicative procedures that are common to Western legal systems. Nevertheless, owing largely to political changes wrought by the end of the Cold War, planning for an African Court of Human Rights was begun in the late 1990s. As envisioned, the Court would not replace the Commission but would supplement and reinforce its mandate.

Four other distinctive features of the African charter are noteworthy. First, it provides for economic, social, and cultural rights as well as civil and political rights. In this respect it resembles the American convention and differs from the European convention. Second, in contrast to both the European and American conventions, it recognizes the rights of groups in addition to the family, women, and children. The aged and the infirm are accorded special protection also, and the right of peoples to self-determination is elaborated in the right to existence, equality, and nondomination. Third, it uniquely embraces the third-generation, or "solidarity," rights to economic, social, and cultural development and to national and international peace and security. Finally, it is to date the only treaty instrument to detail individual duties as well as individual rights—to the family, society, the state, and the international African community. In view of the turmoil that beset northern and sub-Saharan Africa at the end of the 20th century, it is fair to say that the African human rights system is still in its infancy.

Middle Eastern and Asian human rights systems. The Permanent Arab Commission on Human Rights, founded by the Council of the League of Arab States in September 1968, has been occupied primarily with the rights of Arabs living in Israeli-occupied territories. Functioning more to promote than to protect human rights, at the end of the 1990s it had yet to bring a proposed Arab Convention on Human Rights to a successful conclusion. Nevertheless, work by other intergovernmental and nongovernmental bodies manifested a continuing desire to establish human rights protection mechanisms in the Middle East. Building on the Universal Islamic Declaration of Human Rights (1981) and the Cairo Declaration on Human Rights in Islam (1990), the League of Arab States approved an Arab Charter on Human Rights in September 1994. The Charter provides for periodic reports to the league's Human Rights Committee by state parties and for an independent Committee of Experts apparently empowered to request and study reports and submit its own findings to the Human Rights Committee. No other institutions or procedures for monitoring human rights are specified in the Charter, however. More than in most other regions of the world (except Asia), the states of the Middle East were greatly divided over the need to enforce human rights law and the desirability of achieving a true regional system for the promotion and protection of human rights.

The Arab Charter

In Asia, despite efforts by NGOs and the United Nations, the states of the region were at best ambivalent—and at worst hostile—to human rights issues, thus precluding agreement on almost all regional human rights initiatives. In early 1993, anticipating the Vienna World Conference on Human Rights later that year, a conference of Asia-Pacific NGOs adopted an Asia-Pacific Declaration of Human Rights, and in 1997 another meeting of NGOs adopted an Asian Human Rights Charter. Both of these initiatives supported the universality and indivisibility of human rights. However, whereas the first initiative called for the creation of a regional human rights regime, the second—seemingly in deference to the cultural diversity and vastness of the region—urged instead the establishment of national human rights commissions and so-called "People's Tribunals," which would be based more on moral and spiritual foundations rather than on legal ones. The states of Asia were slow to respond to these recommendations. Their positions were indicated at a UN-sponsored workshop in 1996, where the 30 participating states concluded that "it was premature . . . to discuss specific arrangements relating to the setting up of a formal human rights mechanism in the Asian and Pacific region." The same states agreed, however, to "[explore] the options available and the process nec-

The Asian Human Rights Charter

essary for establishing a regional mechanism." It remained to be seen whether the economic and political crises that beset Asia at the end of the 20th century would stimulate regional efforts to ensure greater respect for human rights.

INTERNATIONAL HUMAN RIGHTS IN DOMESTIC COURTS

Using domestic courts to clarify and safeguard international human rights is a new and still evolving approach to human rights advocacy. In addition to the inevitable interpretative problems involved in applying norms that are fashioned in multicultural settings, the approach is also burdened by controversial theories about the interrelation of national and international law and by numerous procedural difficulties. To be sure, considerable progress has been made, as perhaps best evidenced in the far-reaching decision handed down by the U.S. Court of Appeals for the 2nd Circuit in *Filartiga* v. *Pena-Irala* in 1980, in which the court held that the international prohibition of torture must be honoured in U.S. courts. More recently, in 1998–99, the United Kingdom's highest tribunal, the Law Lords of the British House of Lords, captured international attention when, in response to an extradition request by a Spanish court, it upheld the arrest in England of former Chilean President Augusto Pinochet on charges of torture and conspiracy to commit torture in violation of international treaty law. In so doing, it established the precedent that former heads of state do not enjoy immunity from prosecution, at least for systematic human rights crimes.

Human rights at the turn of the 21st century

Whatever the current attitudes and policies of governments, the reality of popular demands for human rights, including both greater economic justice and greater political freedom, is beyond debate. A deepening and widening concern for the promotion and protection of human rights on all fronts, hastened by the self-determinist impulse of a postcolonial era, is now unmistakably woven into the fabric of contemporary world affairs.

Substantially responsible for this progressive development has been the work of the United Nations, its allied agencies, and such regional organizations as the Council of Europe, the Organization of American States, and the Organization of African Unity. Also contributing to this development, particularly since the 1970s and '80s, have been five other salient factors: (1) the public advocacy of human rights as a key aspect of national foreign policies; (2) the emergence and spread of civil society on a transnational basis, primarily in the form of activist nongovernmental human rights organizations such as Amnesty International and Human Rights Watch; (3) a worldwide profusion of teaching and research devoted to the study of human rights; (4) the work of large UN conferences in areas such as children's rights, population, social development, women's rights, human settlements, and food production and distribution; and (5) a mounting intellectual and political challenge from feminists regarding not only the rights of women worldwide but also what they regard as the myths defining humane governance generally.

Contributions of nongovernmental organizations

To be sure, because the application of international human rights law depends for the most part on the voluntary consent of nations, formidable obstacles attend the endeavours of human rights policy makers, activists, and scholars. Human rights conventions continue to be undermined by the failure of states to ratify them and by emasculating reservations and derogations; by self-serving reporting systems that outnumber objective complaint procedures; and by poor financing for the implementation of human rights prescriptions. In short, though a palpable concern for the advancement of human rights is here to stay, the mechanisms for the enforcement of human rights are still in their infancy.

BIBLIOGRAPHY

Human Rights: A Compilation of International Instruments, rev. 5th ed., 2 vol. in 3 (1994–97), published by the United Nations, contains the texts of human rights treaties and other instruments established under the auspices of the United Nations Centre for Human Rights. *Yearbook on Human Rights* (annual), issued by the Secretariat of the United Nations, documents national and international developments in the human rights field. See also BURNS H. WESTON (ed.), *International Law & World Order: Basic Documents*, vols. 1-5 (1994–); FELIX ERMACORA, MANFRED NOWAK, and HANNES TRETTER (eds.), *International Human Rights: Documents and Introductory Notes* (1993); IAN BROWNLIE (ed.), *Basic Documents on Human Rights*, 3rd. ed. (1992); RICHARD B. LILLICH (ed.), *International Human Rights Instruments: A Compilation of Treaties, Agreements, and Declarations of Especial Interest to the United States*, 2nd ed. (1990); ALBERT P. BLAUSTEIN, ROGER S. CLARK, and JAY A. SIGLER (eds.), *Human Rights Sourcebook* (1987); *International Human Rights Instruments of the United Nations, 1948-1982* (1983), published by UNIFO Publishers; and JAMES AVERY JOYCE, *Human Rights: International Documents*, 3 vol. (1978).

Basic works on the subject include BURNS H. WESTON and STEPHEN P. MARKS (eds. and contribs.), *The Future of International Human Rights* (1999); HURST HANNUM, *Guide to International Human Rights Practice*, 3rd ed. (1999); and *Autonomy, Sovereignty, and Self-Determination: The Accommodation of Conflicting Rights*, rev. ed. (1996); S. JAMES ANAYA, *Indigenous Peoples in International Law* (1996); BEVERLY C. EDMONDS and WILLIAM R. FERNEKES, *Children's Rights: A Reference Handbook* (1996); A.S. GEWIRTH, *The Community of Rights* (1996); DAVID GILLIES, *Between Principle and Practice: Human Rights in North-South Relations* (1996); GUY S. GOODWIN-GILL, *The Refugee in International Law*, 2nd ed. (1996); A.H. ROBERTSON and J.G. MERRILLS, *Human Rights in the World*, 4th ed. (1996); HENRY SHUE, *Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy*, 2nd ed. (1996); HENRY J. STEINER and PHILIP ALSTON, *International Human Rights in Context: Law, Politics, Morals* (1996); THOMAS BUERGENTHAL, *International Human Rights in a Nutshell*, 2nd ed. (1995); ASBJØRN EIDE, CATARINA KRAUSE, and ALLAN ROSAS (eds.), *Economic, Social, and Cultural Rights: A Textbook* (1995); UNITED NATIONS CENTRE FOR HUMAN RIGHTS, *United Nations Action in the Field of Human Rights* (1994); ANN ELIZABETH MAYER, *Islam and Human Rights: Tradition and Politics*, 3rd ed. (1999); REBECCA J. COOK (ed.), *Human Rights of Women: National and International Perspectives* (1994); LOUIS HENKIN and JOHN LAWRENCE HARGROVE (eds.), *Human Rights: An Agenda for the Next Century* (1994); IAN MCLLERSON, *International Law, Rights, and Politics: Developments in Eastern Europe and the CIS* (1994); ANDREW CLAPHAM, *Human Rights in the Private Sphere* (1993); JACK DONNELLY, *The Concept of Human Rights* (1985), and *International Human Rights*, 2nd ed. (1998); HURST HANNUM and DANA D. FISCHER (eds.), *U.S. Ratification of the International Covenants on Human Rights* (1993); PHILIP ALSTON (ed.), *The United Nations and Human Rights: A Critical Appraisal* (1992); ABDULLAHI AHMED AN-NA'IM (ed.), *Human Rights in Cross-Cultural Perspectives* (1992); RICHARD PIERRE CLAUDE and BURNS H. WESTON (eds. and contribs.), *Human Rights in the World Community*, 2nd ed. (1992); JAMES CRAWFORD (ed.), *The Rights of Peoples* (1992); ASTRID J.M. DELISSEN and GERARD J. TANJA, *Humanitarian Law of Armed Conflict: Challenges Ahead* (1991); JAMES C. HATHAWAY, *The Law of Refugee Status* (1991); EDWARD LAWSON (compiler), *Encyclopedia of Human Rights*, 2nd ed, (1996); ANTONIO CASSESE, *Human Rights in a Changing World* (1990; originally published in Italian, 1988); BERTRAND BINOCHE, *Critiques des droits de l'homme* (1989); THEODOR MERON, *Human Rights in Internal Strife: Their International Protection* (1987), and *Human Rights Law-Making in the United Nations* (1986); R.J. VINCENT, *Human Rights and International Relations* (1986); THEODOR MERON (ed.), *Human Rights in International Law: Legal and Policy Issues*, 2 vol. (1984, reprinted 1992); DAVID P. FORSYTHE, *Human Rights and World Politics*, 2nd ed., rev. (1989); KAREL VASAK and PHILIP ALSTON (eds.), *The International Dimensions of Human Rights*, 2 vol., trans. from French (1982); RICHARD FALK, *Human Rights and State Sovereignty* (1981); LOUIS HENKIN (ed.), *The International Bill of Rights: The Covenant on Civil and Political Rights* (1981); MYRES S. MCDOUGAL, HAROLD D. LASSWELL, and LUNG-CHU CHEN, *Human Rights and World Public Order: The Basic Policies of an International Law of Human Dignity* (1980); NOAM CHOMSKY and EDWARD S. HERMAN, *The Political Economy of Human Rights*, 2 vol. (1979); ADAMANTIA POLLIS and PETER SCHWAB (eds.), *Human Rights: Cultural and Ideological Perspectives* (1979); JOHN CAREY, *UN Protection of Civil and Political Rights* (1970); RICHARD P. CLAUDE (ed.), *Comparative Human Rights* (1976); MANOUCHEHR GANJI, *The Realization of Economic, Social, and Cultural Rights: Problems, Policies, Progress* (1975); MAURICE CRANSTON, *What Are Human Rights?* (1973); JÓZSEF HALÁSZ (ed.), *Socialist Concept of Human Rights* (1966; originally published in Hungarian, 1965); HERSCH LAUTERPACHT, *International Law and Human Rights* (1950, reissued 1973); RICHARD B. BILDER, "Rethinking International Human Rights: Some Basic Questions," *Wisconsin Law Review*, 171(1): 171–217 (1969); and EGON SCHWELB, *Human Rights and the International Community: The Roots and Growth of the Universal Declaration of Human Rights, 1948–1963* (1964). (B.H.W.)

Humanism

The word *humanism* has been freely applied to a variety of beliefs, methods, and philosophies that place central emphasis on the human realm. Most frequently, however, the term is used with reference to a system of education and mode of inquiry that developed in northern Italy during the 14th century and later spread through Europe and England. Alternately known as "Renaissance humanism," this program was so broadly and profoundly influential that it is one of the chief reasons why the Renaissance is viewed as a distinct historical period. Indeed, though the word *Renaissance* is of more recent coinage, the fundamental idea of that period as one of renewal and reawakening is humanistic in origin. But humanism sought its own philosophical bases in far earlier times and, moreover, continued to exert some of its power long after the end of the Renaissance.

This article is divided into the following sections:

Origin and meaning of the term *humanism* 665
  The ideal of *humanitas* 665
  Other uses 665
Basic principles and attitudes 666
  Classicism 666
  Realism 666
  Critical scrutiny and concern with detail 666
  The emergence of the individual and the idea of the dignity of man 666
  Active virtue 667
Early history 667
The 15th century 669
  Leon Battista Alberti 669
  The Medici and Federico da Montefeltro 669
Later Italian humanism 670
  Things and words 670
  Idealism and the Platonic Academy of Florence 670
  Machiavelli's realism 671
  The achievement of Castiglione 671
  Tasso's Aristotelianism 671
Northern humanism 671
  Desiderius Erasmus 671
  The French humanists 672
  The English humanists 672
Humanism and the visual arts 674
  Realism 674
  Classicism 674
  Anthropocentricity and individualism 674
  Art as philosophy 674
Humanism, art, and science 675
Humanism and Christianity 675
Later fortunes of humanism 676
Conclusion 676
Bibliography 676

ORIGIN AND MEANING OF THE TERM HUMANISM

The ideal of humanitas. The history of the term *humanism* is complex but enlightening. It was first employed (as *humanismus*) by 19th-century German scholars to designate the Renaissance emphasis on classical studies in education. These studies were pursued and endorsed by educators known, as early as the late 15th century, as *umanisti:* that is, professors or students of classical literature. The word *umanisti* derives from the *studia humanitatis,* a course of classical studies that, in the early 15th century, consisted of grammar, poetry, rhetoric, history, and moral philosophy. The *studia humanitatis* were held to be the equivalent of the Greek *paideia.* Their name was itself based on the Latin *humanitas,* an educational and political ideal that was the intellectual basis of the entire movement. Renaissance humanism in all its forms defined itself in its straining toward this ideal. No discussion, therefore, of humanism can have validity without an understanding of *humanitas.*

The *studia humanitatis*

Humanitas meant the development of human virtue, in all its forms, to its fullest extent. The term thus implied not only such qualities as are associated with the modern word *humanity*—understanding, benevolence, compassion, mercy—but also such more aggressive characteristics as fortitude, judgment, prudence, eloquence, and even love of honour. Consequently the possessor of *humanitas* could not be merely a sedentary and isolated philosopher or man of letters but was of necessity a participant in active life. Just as action without insight was held to be aimless and barbaric, insight without action was rejected as barren and imperfect. *Humanitas* called for a fine balance of action and contemplation, a balance born not of compromise but of complementarity. The goal of such fulfilled and balanced virtue was political in the broadest sense of the word. The purview of Renaissance humanism included not only the education of the young but also the guidance of adults (including rulers) via philosophical poetry and strategic rhetoric. It included not only realistic social criticism but also utopian hypotheses, not only painstaking reassessments of history but also bold reshapings of the future. In short, humanism called for the comprehensive reform of culture, the transfiguration of what humanists termed the passive and ignorant society of the "dark" ages into a new order that would reflect and encourage the grandest human potentialities. Humanism had an evangelical dimension. It sought to project *humanitas* from the individual into the state at large.

The role of the classics

The wellspring of *humanitas* was classical literature. Greek and Roman thought, available in a flood of rediscovered or newly translated manuscripts, provided humanism with much of its basic structure and method. For Renaissance humanists, there was nothing dated or outworn about the writings of Plato, Cicero, or Livy. Compared with the typical productions of medieval Christianity, these pagan works had a fresh, radical, almost avant-garde tonality. Indeed, recovering the classics was to humanism tantamount to recovering reality. Classical philosophy, rhetoric, and history were seen as models of proper method—efforts to come to terms, systematically and without preconceptions of any kind, with perceived experience. Moreover, classical thought considered ethics qua ethics, politics qua politics: it lacked the inhibiting dualism occasioned in medieval thought by the often conflicting demands of secularism and Christian spirituality. Classical virtue, in examples of which the literature abounded, was not an abstract essence but a quality that could be tested in the forum or on the battlefield. Finally, classical literature was rich in eloquence. In particular (since humanists were normally better at Latin than they were at Greek) Cicero was considered to be the pattern of refined and copious discourse. In eloquence humanists found far more than an exclusively aesthetic quality. As an effective means of moving leaders or fellow citizens toward one political course or another, eloquence was akin to pure power. Humanists cultivated rhetoric, consequently, as the medium through which all other virtues could be communicated and fulfilled.

Humanism, then, may be accurately defined as that Renaissance movement which had as its central focus the ideal of *humanitas.* The narrower definition of the Italian term *umanisti* notwithstanding, all the Renaissance writers who cultivated *humanitas,* and all their direct "descendants," may be correctly termed humanists.

Other uses. It is small wonder that a term as broadly allusive as *humanism* should be subject to a wide variety of applications. Of these (excepting the historical movement described above) there are three basic types: humanism as classicism, humanism as referring to the modern concept of the humanities, and humanism as human-centredness.

Accepting the notion that Renaissance humanism was

simply a return to the classics, some historians and philologists have reasoned that classical revivals occurring anywhere in history should be called humanistic. St. Augustine, Alcuin, and the scholars of 12th-century Chartres have thus been referred to as humanists. In this sense the term can also be used self-consciously, as in the New Humanism movement in literary criticism led by Irving Babbitt and Paul Elmer More in the early 20th century.

The word *humanities,* which like the word *umanisti* derived from the Latin *studia humanitatis,* is often used to designate the nonscientific scholarly disciplines: language, literature, rhetoric, philosophy, art history, and so forth. Thus it is customary to refer to scholars in these fields as humanists and to their activities as humanistic.

Humanism and related terms are frequently applied to modern doctrines and techniques that are based on the centrality of human experience. In the 20th century the pragmatic humanism of Ferdinand C.S. Schiller, the Christian humanism of Jacques Maritain, and the movement known as secular humanism, though differing from each other significantly in content, all show this anthropocentric emphasis.

Not only is such a large assortment of definitions confusing, but the definitions themselves are often redundant or impertinent. There is no reason to call all classical revivals humanistic when the word classical suffices. To say that professors in the many disciplines known as the humanities are humanists is to compound vagueness with vagueness, for these disciplines have long since ceased to have or even aspire to a common rationale. The definition of humanism as anthropocentricity or human-centredness has a firmer claim to correctness. For obvious reasons, however, it is confusing to apply this word to classical literature.

BASIC PRINCIPLES AND ATTITUDES

Underlying the early expressions of humanism were principles and attitudes that gave the movement a unique character and would shape its future development.

Classicism. Early humanists returned to the classics less with nostalgia or awe than with a sense of deep familiarity, an impression of having been brought newly into contact with expressions of an intrinsic and permanent human reality. Petrarch, the acknowledged founder of the humanistic movement, dramatized his feeling of intimacy with the classics by writing "letters" to Cicero and Livy. Coluccio Salutati remarked with pleasure that possession of a copy of Cicero's letters would make it possible for him to talk with Cicero. Niccolò Machiavelli would later immortalize this experience in a letter that described his own reading habits in ritualistic terms:

The pursuit of reading

> Evenings I return home and enter my study; and at its entrance I take off my everyday clothes, full of mud and dust, and don royal and courtly garments; decorously reattired, I enter into the ancient sessions of ancient men. Received amicably by them, I partake of such food as is mine only and for which I was born. There, without shame, I speak with them and ask them about the reason for their actions; and they in their humanity respond to me.

Machiavelli's term *umanità* ("humanity") means more than kindness; it is a direct translation of the Latin *humanitas.* Machiavelli implies that he shared with the ancients a sovereign wisdom of human affairs. He also describes that theory of reading as an active and even aggressive pursuit that was common among humanists. Possessing a text and understanding its words were not enough; analytic ability and a questioning attitude were necessary before a reader could truly enter the councils of the great. These councils, moreover, were not merely serious and ennobling; they held secrets available only to the astute, secrets the knowledge of which could transform life from a chaotic miscellany into a crucially heroic experience. Classical thought offered insight into the heart of things. In addition, the classics suggested methods by which, once known, human reality could be transformed from an accident of history into an artifact of will. Antiquity was rich in examples, actual or poetic, of epic action, victorious eloquence, and applied understanding. Carefully studied and well employed, classical rhetoric could implement enlightened policy, while classical poetics could carry enlightenment into the very souls of men. In a manner that might seem paradoxical to more modern minds, humanists associated classicism with the future.

Realism. Early humanists shared in large part a realism that rejected traditional assumptions and aimed instead at the objective analysis of perceived experience. To humanism is owed the rise of modern social science, which emerged not as an academic discipline but rather as a practical instrument of social self-inquiry. Humanists avidly read history, taught it to their young, and, perhaps most importantly, wrote it themselves. They were confident that proper historical method, by extending across time their grasp of human reality, would enhance their active role in the present. For Machiavelli, who avowed to treat of men as they were and not as they ought to be, history would become the basis of a new political science. Similarly, direct experience took precedence over traditional wisdom. Leon Battista Alberti's dictum that an essential form of wisdom could be found only "at the public marketplace, in the theatre, and in people's homes" would be echoed by Francesco Guicciardini:

History and experience

> I, for my part, know no greater pleasure than listening to an old man of uncommon prudence speaking of public and political matters that he has not learnt from books of philosophers but from experience and action; for the latter are the only genuine methods of learning anything.

Renaissance realism also involved the unblinking examination of human uncertainty, folly, and immorality. Petrarch's honest investigation of his own doubts and mixed motives is born of the same impulse that led Giovanni Boccaccio in the *Decameron* to conduct an encyclopaedic survey of human vices and disorders. Similarly critical treatments of society from a humanistic perspective would be produced later by Erasmus, More, Castiglione, Rabelais, and Montaigne. But it was typical of humanism that this moral criticism did not, conversely, postulate an ideal of absolute purity. Humanists asserted the dignity of normal earthly activities and even endorsed the pursuit of fame and the acquisition of wealth. The emphasis on a mature and healthy balance between mind and body, first implicit in Boccaccio, is evident in the work of Giannozzo Manetti, Francesco Filelfo, and Paracelsus; it is embodied eloquently in Montaigne's final essay, "Of Experience." Humanistic tradition, rather than revolutionary inspiration, would lead Francis Bacon to assert in the early 17th century that the passions should become objects of systematic investigation. The realism of the humanists was, finally, brought to bear on the Roman Catholic Church, which they called into question not as a theological structure but as a political institution. Here as elsewhere, however, the intention was neither radical nor destructive. Humanism did not aim to remake humanity but rather to reform social order through an understanding of what was basically and inalienably human.

Critical scrutiny and concern with detail. Humanistic realism bespoke a comprehensively critical attitude. Indeed, the productions of early humanism constituted a manifesto of independence, at least in the secular world, from all preconceptions and all inherited programs. The same critical self-reliance shown by Coluccio Salutati in his textual emendations and Boccaccio in his interpretations of myth was evident in almost the whole range of humanistic endeavour. It was cognate with a new specificity, a profound concern with the precise details of perceived phenomena, that took hold across the arts and the literary and historical disciplines and would have profound effects on the rise of modern science. The increasing prominence of mathematics as an artistic principle and academic discipline was a testament to this development.

The emergence of the individual and the idea of the dignity of man. These attitudes took shape in concord with a sense of personal autonomy that first was evident in Petrarch and later came to characterize humanism as a whole. An intelligence capable of critical scrutiny and self-inquiry was by definition a free intelligence; the intellectual virtue that could analyze experience was an integral part of that more extensive virtue that could, according to many humanists, go far in conquering fortune. The

emergence of Renaissance individualism was not without its darker aspects. Petrarch and Alberti were alert to the sense of estrangement that accompanies intellectual and moral autonomy, while Machiavelli would depict, in *The Prince*, a grim world in which the individual must exploit the weakness of the crowd or fall victim to its indignities. But happy or sad, the experience of the individual had taken on a heroic tone. Parallel with individualism arose, as a favourite humanistic theme, the idea of the dignity of man. Backed by medieval sources but more sweeping and insistent in their approach, spokesmen such as Petrarch, Manetti, Valla, and Ficino asserted man's earthly preeminence and unique potentialities. In his noted *De hominis dignitate oratio* ("Oration on the Dignity of Man"), Giovanni Pico della Mirandola conveyed this notion with unprecedented vigour. Humanity, Pico asserted, had been assigned no fixed character or limit by God but instead was free to seek its own level and create its own future. No dignity, not even divinity itself, was forbidden to human aspiration. Pico's radical affirmation of human capacity shows the influence of Ficino's recent translations of the Hermetic writings. Together with the even bolder 16th-century formulations of this position by Paracelsus and Giordano Bruno, the *Oratio* betrays a rejection of the early humanists' emphasis on balance and moderation; it suggests the straining toward absolutes that would characterize major elements of later humanism.

Active virtue. The emphasis on virtuous action as the goal of learning was a founding principle of humanism and (though sometimes sharply challenged) continued to exert a strong influence throughout the course of the movement. Salutati, the learned chancellor of Florence whose words could batter cities, represented in word and deed the humanistic ideal of an armed wisdom: that combination of philosophical understanding and powerful rhetoric which alone could effect virtuous policy and reconcile the rival claims of action and contemplation. In *De ingenuis moribus et liberalibus studiis* ("On the Manners of a Gentleman and Liberal Studies"), a treatise that influenced Guarino Veronese and Vittorino da Feltre, Pietro Paolo Vergerio maintained that just and beneficent action was the purpose of humanistic education; his words were echoed by Alberti in *Della famiglia* ("On the Family"):

Armed wisdom

> As I have said, happiness cannot be gained without good works and just and righteous deeds. . . . The best works are those that benefit many people. Those are most virtuous, perhaps, that cannot be pursued without strength and nobility. We must give ourselves to manly effort, then, and follow the noblest pursuits.

Matteo Palmieri wrote that

> the true merit of virtue lies in effective action, and effective action is impossible without the faculties that are necessary for it. He who has nothing to give cannot be generous. And he who loves solitude can be neither just, nor strong, nor experienced in those things that are of importance in government and in the affairs of the majority.

Palmieri's philosophical poem, *La città di vita* ("The City of Life"), developed the idea that the world was divinely ordained to test human virtue in action. Later humanism would broaden and diversify the theme of active virtue. Machiavelli saw action not only as the goal of virtue but also (via historical understanding of great deeds of the past) as the basis for wisdom. Baldassare Castiglione, in his highly influential *Libro del cortegiano* (*Book of the Courtier*), developed in his ideal courtier a psychological model for active virtue, stressing moral awareness as a key element in just action. François Rabelais used the idea of active virtue as the basis for anticlerical satire. In his profusely humanistic *Gargantua*, he has the active hero Friar John save a monastery from enemy attack, while the monks sit uselessly in the church choir, chanting meaningless Latin syllables. John later asserts that, had he been present, he would have used his manly strength to save Jesus from crucifixion, and he castigates the Apostles for betraying Christ "after a good meal." Endorsements of active virtue, as will be shown, would also characterize the work of English humanists from Sir Thomas Elyot to John Milton. They typify the sense of social responsibility, the instinctive association of learning with politics and morality, that stood at the heart of the movement. As Salutati put it, "One must stand in the line of battle, engage in close combat, struggle for justice, for truth, for honour."

Rabelais's *Gargantua*

EARLY HISTORY

The influence of Petrarch (Francesco Petrarca, 1304–74) was profound and many-sided. As the most prominent man of letters of the 14th century, he promoted the recovery and transcription of classical texts, providing the impetus for the important classical researches of Boccaccio and Salutati. He threw himself into controversies in which he defined a new humanism in contradistinction to what he considered to be the barbaric influence of medieval tradition. He carried on an energetic correspondence that established him as a cultural focal point and would provide, if all his other works were lost, an accurate index of his views and their development. As a theologian (he was an ordained priest) he advanced the view, held by many humanists to follow, that classical learning and Christian spirituality were not only compatible but also mutually fulfilling. As a political apologist, he gave hearty support to Cola di Rienzo's brief revival of the Roman Republic (1347). As a poet, he was the first Renaissance writer to produce a Latin epic (*Africa*), but he was even more important for his compositions in the vernacular. His *Canzoniere* provided the model on which the Renaissance lyric was to take shape and the standard by which future productions would be judged. His work established secular poetry as a serious and noble pursuit. His eloquent and forceful presence made him a personal symbol of his own ideas. Crowned with laurel, favoured by rulers, legates, and scholars, he became the human focus for the new interest in classical revival and literary artistry.

Petrarch's greatest influence

It was, however, as a philosophical spokesman that Petrarch exerted his greatest influence on the history of humanism. In his prose works and letters he established many of the positions that would be central to the movement and broached many of the issues that would be its favourite subjects for debate. His idea of the poet as a philosophical teacher and thus as a champion of culture would inspire humanists from Boccaccio to Sidney. His endorsement of the study of rhetoric and his underlying notion of language as an informing principle of the individual and society would become crucial subjects of humanistic discussion and debate. His view of classical culture, not as an undifferentiated element of the past but as an authentic alternative to his own medieval society, was of equal historical importance. Petrarch broke with the past and helped to reestablish the Socratic tradition in Europe by specifying self-knowledge as a primary goal of philosophy. This attitude and his unfailing insistence on moral autonomy were early and important signs of the individualism that would become a Renaissance hallmark. He emphasized human virtue as opposed to fortune, thus setting the stage for numerous famous treatments of this theme. He struggled repeatedly with the dilemma of action versus contemplation, establishing it as a favourite topic for humanistic debate. Petrarch did not invent these subjects, nor does he usually treat them with overwhelming power. His preeminence lies in the fact that he was the first writer since antiquity to assert that they and other human matters were valid issues for philosophical inquiry in and of themselves, and in the energy and eloquence with which he made his work their forum.

Petrarch's influence was immediately apparent in the work of two major Florentine humanists, Giovanni Boccaccio and Coluccio Salutati. A close friend and devoted supporter of Petrarch, Boccaccio (1313–75) not only enlarged upon his preceptor's ideas but also made important humanistic contributions of his own. His *Teseide* was the first classical epic to have been written in the vernacular and influenced the more famous Italian epics of Ariosto and Tasso. His *De genealogia deorum gentilium* ("On the Genealogy of the Gods of the Gentiles"), a scholarly interpretive compendium of classical myth, was the first in a long line of Renaissance mythographies; it includes a celebrated defense of poetry as a medium of hidden truth, a stimulant to virtue, and a source of mental health. His most memorable contribution to humanism, however, was

Boccaccio's *Decameron*

probably the famous *Decameron.* Ostensibly this work is no more than a collection of 100 tales about love. But subjected to the interpretive scrutiny that Boccaccio himself recommends in *De genealogia deorum gentilium,* the *Decameron* takes on a far more serious tone. The opening phrase "*Umana cosa è*" ("It is a human thing") is deeply thematic, reminding us that the author structured his work on Dante's spiritual epic, *La divina commedia.* A close reading of the *Decameron* suggests that in it Boccaccio is trying to establish for the human realm the same sort of comprehensive understanding that Dante established for the life of the spirit. Through moral fable and direct address to the reader, he undertakes a reinterpretation of human experience based not on traditional doctrine but rather on perceived reality. Appealing repeatedly to reason and nature, and constantly implying the superiority of awareness to innocence (which he equates with ignorance), he calls for a moral order built fairly and solidly on the potentialities of human nature. His 10 storytellers, who leave the plague-ravaged and chaotic city of Florence and reestablish themselves at a delightfully landscaped villa, suggest the remaking of culture through disentanglement with the past, unprejudiced analysis, and enlightened imagination. Rightly considered to be the wellspring of Western realism, the *Decameron* is also a monument to humanism. Though it makes little mention of classical thought, Boccaccio's great work rings with a tone that was even more basic to the humanistic movement: an emphasis on the human capacity for self-knowledge and willed renewal.

Coluccio Salutati

Other humanistic elements implicit in Petrarch's thought were developed in the life and work of Coluccio Salutati (1331–1406). Like Petrarch, Salutati collected manuscripts, wrote on morality and politics, and carried on a voluminous correspondence. He was an aggressive and scientific philologist, instrumental in establishing principles of textual criticism that would become key elements of the humanistic method. He was a forceful apologist for the active life, and his theories bore fruit in his own career as chancellor of the Florentine republic. His use of classical eloquence in the service of his state was an early documentation of the humanistic faith in the political power of rhetoric; it led a bitter enemy, Gian Galeazzo Visconti of Milan, to say that a thousand Florentine horsemen had hurt him less than the letters of Coluccio. Salutati was succeeded in the Florentine chancellorship by two scholar-statesmen who reflected his influence, first Leonardo Bruni (1369–1444) and then Gian Francesco Poggio Bracciolini (1380–1459). Bruni was a pioneer in the advocacy of humanistic education, holding that the *studia humanitatis* shape the perfected man and that the goal of this perfected virtue is political action. His theory of education stressed the importance of practical experience (implicit in the work of Boccaccio) and put heavy emphasis on historical studies. His history of Florence is considered to be the first work of modern historiography; and, under the influence of Emmanuel Chrysoloras (1368–1415), a Byzantine teacher who had lectured at Florence and Pavia, he produced Latin translations of Plato and Aristotle that broke with medieval tradition by reproducing the sense of the Greek prose rather than following it word by word. Poggio, the foremost recoverer of classical texts, was also a moralist, a historian, a brilliant correspondent, and an early scholar of architectural antiquities. His long career, which included service to both church and state and friendships with Salutati, Bruni, Niccolò Niccoli, Guarino, Nicholas of Cusa, Donatello, and Cosimo de' Medici, exemplifies the scope and vitality of Italian humanism. Together these Florentine chancellors, whose active lives spanned almost a century, strengthened and consolidated the humanistic program. Moreover, their leadership strongly influenced the cultural developments that would make 15th-century Florence the most active intellectual and artistic centre in Europe.

As one proceeds with the history of humanism, the following major points about its development in the 14th century ought to be kept in mind. Humanism received its crucial imprint from the work of a single man and thence developed among men who maintained close touch with each other and acknowledged a shared mission. Humanism was not originally an academic movement but rather a program defined and promoted by statesmen and men of letters. Its proclaimed goal was widespread cultural renewal; therefore, it chose its subjects for consideration from the phenomena of human life as lived and adopted the Ciceronian model of philosopher as citizen in preference to the contemplative ideal. The heavy emphasis on civic action is connected with the fact that humanism developed in a republic rather than a monarchy.

The humanistic schools

By the turn of the 15th century, all of the key elements that came to define humanism were in place except for two: its detailed educational system and what might be called its Greek dimension. The founders of the first humanistic schools were Vittorino da Feltre (1373–1446) and Guarino Veronese (Guarino da Verona, 1374–1460). Vittorino and Guarino were fellow students at the University of Padua at the turn of the century; they are said later to have tutored each other (Guarino as an expert in Greek, Vittorino in Latin) after Guarino had opened the first humanistic school (Venice, *c.* 1414). Vittorino taught in both Padua (where he was briefly professor of rhetoric) and Venice during the early 1420s. In 1423 he accepted the invitation of Gianfrancesco Gonzaga, marquis of Mantua, to become tutor to the ruling family. At this post Vittorino spent the remaining 22 years of his life. His school, held in a delightful palace that he renamed "La Giocosa," had as its students not only the Gonzaga children (among them the future marquis, Ludovico) but also an increasing number of others, including sons of Poggio, Guarino, and Filelfo. The eminent humanist Lorenzo Valla studied there, as did Federico da Montefeltro, who later promoted humanistic institutions as duke of Urbino. Vittorino's school in Mantua was the first to focus the full power of the humanistic program, together with its implications in other arts and sciences, upon the education of the young. Latin literature, Latin composition, and Greek literature were required subjects of study. Heavy emphasis was placed on Roman history as an educational treasury of great men and memorable deeds. Rhetoric (as taught by Quintilian) was a central topic, not as an end in itself but as an effective means of channeling moral virtue into political action. Vittorino summed up the essentially political thrust of humanistic education as follows:

> Not everyone is called to be a physician, a lawyer, a philosopher, to live in the public eye, nor has everyone outstanding gifts of natural capacity, but all of us are created for the life of social duty, all are responsible for the personal influence that goes forth from us.

Other studies at Mantua included music, drawing, astronomy, and mathematics. The meadows around La Giocosa were turned into playing fields. Vittorino's educational policy spoke at once to mind and body, to aesthetic enjoyment and moral virtue. His work embodied a more comprehensive appeal to human perfectibility than had been attempted since antiquity. Humanists were not unaware of the originality and ambitiousness of this project. With reference to a similar program of his own, Guarino's son Battista remarked that "no branch of knowledge embraces so wide a range of subjects as that learning that I have now attempted to describe."

Byzantine influence

Guarino had learned his Greek in Constantinople under the influence of Chrysoloras, whose dynamic presence had done much to foster Greek studies in Italy. During the course of the 15th century, which saw the famous council of Eastern and Western churches (Ferrara–Florence, 1438–45) and later the fall of Constantinople to the Turks (1453), Italy received as welcome immigrants a number of other eminent Byzantine scholars. George Gemistus Plethon (1355–1450) was a major force in Cosimo de' Medici's foundation of the Platonic Academy of Florence. George of Trebizond (Georgius Trapezuntius, 1395–1484), a student of Vittorino, was a formidable bilingual stylist who wrote important handbooks on logic and rhetoric. Theodore Gaza (*c.* 1400–75) and Johannes Argyropoulos (1410–90) contributed major translations of Aristotle. John (originally Basil) Bessarion (1403–72), who became a cardinal in 1439, explored theology from a Platonic perspective and sought to resolve apparent conflicts between Platonic and Aristotelian philosophy; his large collection

of Greek manuscripts, donated to the Venetian senate, became the core of the notable library of St. Mark. This infusion of Byzantine scholarship had a profound effect on Italian humanism. By making Greek texts and commentaries available to Western students, and by acquainting them with Byzantine methods of criticism and interpretation, the teachers from Constantinople enabled Italian humanists to explore the bases of classical thought and to appreciate its greatest monuments, either in the original or in accurate new Latin translations.

THE 15TH CENTURY

As Italian humanism grew in influence during the 15th century, it developed ramifications that connected it with every major field of intellectual and artistic activity. Moreover, the advent of printing at mid-century and the contemporaneous upsurge of publication in the vernacular brought new sectors of society under humanistic influence. These and other cultural impetuses hastened the export of humanistic ideas to the Low Countries, France, England, and Spain, where significant humanistic programs would be in place by the early 16th century. Even as these things were happening, however, other changes were deeply and permanently affecting the character of the movement. The concerns of many major humanists were narrowed by inevitable historical processes of specialization, to the extent that, in a large number of cases, humanism lost its comprehensive thrust and became a predominantly academic or literary pursuit. The political élan of humanism was weakened by the decline of republican institutions in Florence. Ambiguities and paradoxes implicit in the original program developed into open conflicts, dividing the movement into camps and depleting much of its original integrity. But before considering these developments, one might do well to appreciate three 15th-century examples of humanism at its height: the career of Leon Battista Alberti and the humanistic courts at Florence and Urbino.

Leon Battista Alberti. The achievement of Leon Battista Alberti (1404–72) testifies to the formative power and exhaustive scope of earlier Italian humanism. He owed his boyhood education to Gasparino da Barzizza (1359–1431), the noted teacher who, with Vergerio, was influential in the development of humanism at Padua. Alberti attended the University of Bologna from 1421 until 1428, by which time he was expert in law and mathematics and so adept at humanistic literary skills that his comedy *Philodoxeos* was accepted as the newly discovered work of an ancient author. In 1428 he became secretary to Cardinal Albergati, bishop of Bologna, and in 1432 he accepted a similar position in the papal chancery at Rome. His service to the church soon brought him incomes that permanently secured his livelihood, and he spent the remainder of his life at a variety of literary, philosophical, and artistic pursuits so dazzling as to challenge belief. He was a poet, essayist, and biographer. His moral and philosophical works, especially *Della famiglia, De iciarchia* ("On the Man of Excellence and Ruler of His Family"), and *Momus,* are humanistic statements that nonetheless bear the mark of a unique individual. He wrote a rhetorical handbook and a grammatical treatise, the *Regule lingue Florentine,* which bespeaks his strong influence on the rise of literary expression in the vernacular. He contributed an important text on cartography and was instrumental in the development of ciphers. A prominent architect (*e.g.,* the Tempio Malatestiano in Rimini and the facade of Sta. Maria Novella in Florence), he was also an eminent student of all artistic ideas and practices. His three studies—*De pictura* (*On Painting*), *De statua* (*On Sculpture*), and *De re aedificatoria* (*Ten Books on Architecture*)—were landmarks in art theory, powerful in developing the theory of perspective and the idea of "human" space. His theoretical and practical reliance on mathematics (which he considered to be the basic, unifying element of all science) is rightly seen as an important step in the early development of modern method.

Alberti's writings

Behind these achievements was a man of startling physical prowess and inexhaustible sanguinity. He said outright that an individual could encompass whatever project he truly willed, and his own life bore witness to this radical thesis. In the 19th century Jacob Burckhardt would write of him as a "universal man" of the Renaissance, while his own contemporary Politian described him with wonderment: "It is better to be silent about him than not to say enough." Alberti's theory and practice bore an undeniably humanistic stamp. His passion for mathematics was in all likelihood an outgrowth of the educational program at Padua (Vittorino, himself an avid mathematician, was also a student of Barzizza). His omnivorous pursuit of knowledge recalls Barzizza's conviction that *humanitas* was the unifying principle of many arts. An advocate of classical erudition in art and architecture as well as in literary activity, he extended into his artistic studies the same sense of precision and specificity that earlier humanists had applied to philology. His sense of human dignity, evident in all his productions, was supported and indeed justified by a strenuous realism. His advocacy of the vernacular disturbed a number of more doctrinaire humanists, who favoured total Latinity. But this predisposition, rather than a divergence from humanistic principle, was a direct outgrowth of its evangelistic thrust. In short, Alberti uniquely fulfilled the humanistic aspiration for a learning that would comprehend all experience and a philosophical heroism that would renew society.

The Medici and Federico da Montefeltro. The 15th century saw the rise of the Platonic Academy of Florence and the great humanistic courts. Close ties between Poggio and the Medici helped make that ruling family of Florence the new custodians of the humanistic heritage. Cosimo de' Medici (Cosimo the Elder, 1389–1464), who had personally lured the great council of churches from Ferrara to Florence in 1439, became so enamoured of Greek learning that, at the suggestion of Gemistus Plethon, he decided to found a Platonic academy of his own. He amassed a great collection of books, which would form the nucleus of the Laurentian Library. He generously supported the work of scholars, in particular encouraging the brilliant Marsilio Ficino (1433–99) to undertake a complete Latin translation of Plato. Other notable members of the academy were Politian, Cristoforo Landino (1424–1504), and Ficino's own student, Giovanni Pico della Mirandola (1463–94). The Medici family was equally notable in its patronage of the arts, supporting projects by a list of masters that included Brunelleschi, Michelangelo, and Cellini. Cosimo's famous grandson Lorenzo (Lorenzo the Magnificent, 1449–92) was of a thoroughly humanistic disposition. Lorenzo's versatile and energetic nature lent itself equally to politics and philosophy, to martial arts and music. He wrote poetry and literary commentary and formed close ties with Ficino, Pico, and other leading scholars of the academy. He continued his grandfather's lavish patronage of art and learning and was said to have spent half of his city's revenues on the purchase of books alone. Active in many fields, he nonetheless acknowledged the preeminence of the life of the mind. When chided by a friend for sleeping late and not going out to work, Lorenzo replied, "What I have dreamed in one hour is worth more than what you have done in four."

The Medici patronage

The influence of humanism was evident in many 15th-century Italian courts, including Rome itself, which boasted, in Pius II (Enea Silvio Piccolomini, also known as Aeneas Sylvius Piccolomini, 1405–64), a humanist pope. It manifested itself strikingly at Urbino, where Federico da Montefeltro (1422–82) turned an isolated hill town into a treasury of Renaissance culture. Schooled by Vittorino in Mantua, Federico chose warfare as his calling. As a mercenary he gained a reputation for winning his battles and keeping his word, and the fortune he accumulated in fees and prizes became the medium for his city's renewal. He brought architects, artists, and scholars to Urbino and built a great palace whose unadorned exterior concealed magnificent chambers, a graceful courtyard, and a secret garden. Federico was enthusiastically devoted to the collection and preservation of books. His library, described by Vespasiano Bisticci as being even more complete than that of the Medici, contained an army of 30 to 40 scribes who were constantly at work. His own virtues were so notable and diverse as to mark him as a possible model for Rabelais's humanistic giant, Gargantua. Mighty at

Federico's library

arms, he was also conscientious in religious observances; supremely powerful, he was nonetheless a modest and courteous companion. Beneath the ivied tranquility of his secret garden stretched an indoor equestrian arena. He commissioned paintings by Piero della Francesca and was the object of humanistic dedications by Poggio, Landino, and Ficino. He kept two organists at court and maintained five men to read the classics aloud at meals. Federico's intellectual accomplishments were impressive. His skill at mathematics shows the influence of Vittorino. He was a good Latinist and as a student of classical history was able to hold his own in conversation with the erudite Pius II. At philosophy Federico was even more astute. Vespasiano wrote that

> he began to study logic with the keenest understanding, and he argued with the most nimble wit that was ever seen. After he had heard (Aristotle's) *Ethics* many times, comprehending it so thoroughly that his teachers found him hard to cope with in disputation, he studied the *Politics* assiduously. . . . Indeed, it may be said of him that he was the first of the Signori who took up philosophy and had knowledge of the same. He was ever careful to keep intellect and virtue to the front, and to learn some new thing every day.

Federico's balance and versatility made him, even more than Lorenzo, an example of the humanistic program in action. Baldassare Castiglione, perhaps the most thoughtful of the later Italian humanists, would speak of him as "the light of Italy; there is no lack of living witnesses to his prudence, humanity (*umanità*), justice, intrepid spirit, (and) military discipline." Castiglione described Federico's residence as seeming to be less a palace than "a city in the form of a palace"; one might say as well that this structure, with its elegant accommodation for every creative human activity, was an architectural image of the humanistic mind.

LATER ITALIAN HUMANISM

The achievement of Alberti, Federico, and the Medici up to Lorenzo may be seen as the effective culmination of Italian humanism, the ultimate realization of its motives and principles. At the same time as these goals were being achieved, however, the movement was beginning to suffer bifurcation and dilution. Even the enthusiastic Platonism of the Florentine academy was, in its idealism and emphasis on contemplation, a significant digression from the crucial humanistic doctrine of active virtue, and Pico della Mirandola himself was politely admonished by a friend to forsake the ivory tower and accept his civic responsibilities. The conflicting extremes to which sincere humanistic inquiry could drive scholars are nowhere more apparent than in the fact that the arch-idealist Pico and the arch-realist Machiavelli lived in the same town and at the same time. Castiglione, who had belonged to the court of Federico's son Guidobaldo, would be saddened by its decline and shocked when another of his patrons, the "model" Renaissance prince Charles V, ordered the sack of Rome. To a large extent, the cause of these and other vicissitudes lay in the nature of the movement itself, for that boundless diversity which nourished its strength was also a well of potential conflict. Humanists' undifferentiated acceptance of the classical heritage was also in effect an appropriation of the profound controversy implicit in that heritage. Rifts between Platonists, monarchists, and republicans; positivists and skeptics; idealists and cynics; and historians and poets came to be more and more characteristic of humanistic discourse. Some of these tensions had been clear from the start, Petrarch having been ambiguous in his sentiments regarding action versus contemplation, and Salutati having been not wholly clear about whether he preferred republics to monarchies. But the 15th century, bringing with it the irreconcilable heterogeneity of Greek thought, vastly multiplied and deepened these divisions. Of these schisms, the two that perhaps most deeply influenced the course of humanism were the so-called *res–verbum* ("thing–word") controversy and the split between Platonic idealism and historical realism.

Diversification and erosion

Things and words. Simply put, the *res–verbum* controversy was an extended argument between humanists who believed that language constituted the ultimate human reality and those who believed that language, though an important subject for study, was the medium for understanding an even more basic reality that lay beyond it. The origin of the controversy lay in the debate in the 5th–4th century BC between the Socratic school, which held that language was an important means of understanding deeper truths, and the Sophistic-rhetorical school, which held that "truth" was itself a fiction dependent on varying human beliefs and therefore that language had to be considered the ultimate arbiter. Petrarch, who had no direct contact with the works of Plato and little detailed knowledge of his ideas, drew on Cicero and St. Augustine in his development of a Christian-rhetorical position, holding that "it is more satisfying (*satius*) to will the good than to know the truth" and espousing rhetoric as the effective means of convincing people "to will the good."

Origin of the *res–verbum* controversy

This assertion would critically shape the character of humanism through the Renaissance and beyond. It was never effectively challenged by Renaissance Platonists because, for reasons discussed below, Renaissance Platonists, though strong in Platonic idealism, were weak in Platonic analytical method. The enthronement of language as both subject and object of humanistic inquiry is evident in the important work of Lorenzo Valla (1407–57) and Politian (Angelo Poliziano, 1454–94). Valla spoke of language as a "sacrament" and urged that it be studied scientifically and historically as the synthesis of all human thought. For Valla, the study of language was, in effect, the study of humanity. Similarly, Politian held that there were in fact two dialectics: one of ideas and one of words. Rejecting the dialectic of ideas as being too difficult and abstruse, he espoused the dialectic of words (*i.e.*, philology and rhetoric) as the proper human study. This project would bear fruit in the intensive linguistic-philosophical researches of Mario Nizolio (1498–1575). Though anticipated by Petrarch, the radical emphasis on the primacy of the word constituted a break with the teaching of other early humanists, such as Bruni and Vittorino, who had strongly maintained that the word was of value only through its relationship to perceived reality. Nor did the old viewpoint lack later adherents. In an epistolary debate with Ermolao Barbaro (1454–93), Pico asserted the preeminence of things over words and hence of philosophy over rhetoric: "But if the rightness of names depends on the nature of things, is it the rhetorician we ought to consult about this rightness, or is it the philosopher who alone contemplates and explores the nature of everything?" Appeals of this sort, however, were not to win the day. Philosophical humanism declined because, though rich in conviction, it had failed to establish a systematic relationship between philosophy and rhetoric, between words and things. By the 16th century, Italian humanism was primarily a literary pursuit, and philosophy was left to develop on its own. Despite significant challenges, the division between philosophical and literary studies would solidify in the development of Western culture.

Idealism and the Platonic Academy of Florence. The idealism so prominent in the Florentine academy is called Platonic because of its debt to Plato's theory of Ideas and to the epistemological doctrine established in his *Symposium* and *Republic*. It did not, however, constitute a complete appreciation or reassertion of Plato's thought. Conspicuously absent from the Florentine agenda was the analytic method (dialectic), which was Socrates' greatest contribution to philosophy. This major omission cannot be explained philologically, at least after Ficino's work had made the complete Platonic corpus available in clear Latin prose. The explanation lies rather in a specific cast of mind and in a dramatically successful forgery. The major Platonists of the mid-15th century, Plethon, Bessarion, and Nicholas of Cusa (Nicholaus Cusanus, 1401–64), had all concentrated their attention on the religious implications of Platonic thought; and, following them, Marsilio Ficino (1433–99) sought to reconcile Plato with Christ in a *pia philosophia* ("pious philosophy"). The transcendental goals of these philosophers left little room for the painstaking dialectical method that sifted through the details of perception and language, even though Plato himself had repeatedly alleged that transcendence itself was impos-

sible without this method. Along with Plato, moreover, Ficino had translated into Latin the works of the so-called Hermes Trismegistos. These books, which also emphasized transcendence at the expense of method, laid claim to divine authority and to an antiquity far greater than Plato's. They were, in fact, forgeries from a much later period, and are in many ways typical of the idealized and diluted versions of Plato that are called Neoplatonic. But the academy, and for that matter all the other Platonists of the 15th century, bought them wholesale. The result of these factors was a Platonism sans Platonic method, a philosophy that, straining for absolutes, had little interest in establishing its own basis in reality. Near the end of *The Book of the Courtier,* Castiglione puts a speech typical of Florentine Platonism in the mouth of his friend, the Platonist Pietro Bembo (1470–1547). As Bembo finishes his oration, a female companion tugs at the hem of his robe and says, "Take care, Master Pietro, that with such thoughts your soul does not forsake your body."

Machiavelli's realism. Niccolò Machiavelli (1469–1527), whose work derived from sources as authentically humanistic as those of Ficino, proceeded along a wholly opposite course. A throwback to the chancellor-humanists Salutati, Bruni, and Poggio, he served Florence in a similar capacity and with equal fidelity, using his erudition and eloquence in a civic cause. Like Vittorino and other early humanists, he believed in the centrality of historical studies, and he performed a signally humanistic function by creating, in *La Mandragola,* the first vernacular imitation of Roman comedy. His characteristic reminders of human weakness suggest the influence of Boccaccio; and like Boccaccio he used these reminders less as satire than as practical gauges of human nature. In one way at least, Machiavelli is more humanistic (*i.e.,* closer to the classics) than the other humanists, for while Vittorino and his school ransacked history for examples of virtue, Machiavelli (true to the spirit of Polybius, Livy, Plutarch, and Tacitus) embraced all of history, good, evil, and indifferent, as his school of reality. Like Salutati, though perhaps with greater self-awareness, Machiavelli was ambiguous as to the relative merits of republics and monarchies. In both public and private writings (especially the *Discorsi sopra la prima deca di Tito Livio* ["Discourses on the First Ten Books of Livy"]) he showed a marked preference for republican government, while in *The Prince* he developed, with apparent approval, a model of radical autocracy. For this reason, his goals have remained unclear.

Machiavelli's writings

His methods, on the other hand, were coherent throughout and remain a major contribution to social science and the history of ideas. Like earlier humanists, Machiavelli saw history as a source of power, but, unlike them (and here perhaps influenced by Sophistic and Averroistic thought), he saw neither history nor power itself within a moral context. Rather he sought to examine history and power in an amoral and hence (to him) wholly scientific manner. He examined human events in the same way that Alberti, Galileo, and the new science examined physical events: as discrete phenomena that had to be measured and described before they could be explained and evaluated. To this extent his work, though original in its specific design, was firmly based in the humanistic tradition. At the same time, however, Machiavelli's achievement significantly eroded humanism. By laying the foundations of modern social science, he created a discipline that, though true to humanistic methodology, had not the slightest regard for humanistic morality. In so doing, he brought to the surface a contradiction that had been implicit in humanism all along: the dichotomy between critical objectivity and moral evangelism.

The achievement of Castiglione. Though Italian humanism was being torn apart by the natural development of its own basic motives, it did not thereby lose its native attractions. The humanistic experience, in both its positive and negative effects, would be reenacted abroad. Baldassare Castiglione (1478–1529), whose *Book of the Courtier* affectionately summed up humanistic thought, was one of its most powerful ambassadors. Alert to the major contradictions of the program, yet intensely appreciative of its brilliance and energy, Castiglione wove its various strains together in a long dialogue that aimed at an equipoise between various humanistic extremes. Ostensibly a treatise on the model courtier, *The Book of the Courtier* is more seriously a philosophically organized pattern of conflicting viewpoints in which various positions—Platonist and Aristotelian, idealist and cynic, monarchist and republican, traditional and revolutionary—are given eloquent expression. Unlike most of his humanistic forebears, Castiglione is neither missionary nor polemical. His work is not an effort at systematic knowledge but rather an essay in higher discretion, a powerful reminder that every virtue (moral or intellectual) suggests a concomitant weakness and that extreme postures tend to generate their own opposites. The structure of the dialogue, in which Bembo's Platonic ecstasy is balanced by Bibbiena's assortment of earthy jests, is a testament to this intention. While Castiglione's professed subject matter would epidemically inspire European letters and manners of the 16th century, his more profound contribution would be echoed in the work of Montaigne and Shakespeare. His work suggests a redefined humanism, a virtue matured in irony and directed less toward knowledge than toward wisdom.

The Book of the Courtier

Tasso's Aristotelianism. In 16th-century Italy, humanistic methods and attitudes provided the medium for a kaleidoscopic variety of literary and philosophical productions. Of these, the work that perhaps most truly reflected the original spirit of humanism was the *Gerusalemme liberata* of Torquato Tasso (1544–95). New humanistic translations of Aristotle during the 15th century had inspired an Aristotelian Renaissance, and the attention of literary scholars focused particularly on the *Poetics.* In constructing his epic poem, Tasso was strongly influenced by Aristotle's views regarding the philosophical dimension of poetry; loosely paraphrasing Aristotle, he held (in his *Apologia*) that poetry, by incorporating both particulars and universals, was capable of seeking truth in its perfect wholeness. As a vehicle for philosophical truth, poetry consequently could provide moral education, specifically in such virtues (reinterpreted from a Christian perspective) as Aristotle had described in the *Nichomachean Ethics.* The Aristotelian Renaissance thus facilitated the revival of one of the chief articles in the original humanistic constitution: the belief in the poet's role as renewer of culture.

NORTHERN HUMANISM

Though humanism in northern Europe and England sprang largely from Italian sources, it did not emerge exclusively as an outgrowth of later Italian humanism. Non-Italian scholars and poets found inspiration in the full sweep of the Italian tradition, choosing their sources from Petrarch to Castiglione and beyond.

Desiderius Erasmus. Erasmus (*c.* 1466–1536) was the only other humanist whose international fame in his own time compared with Petrarch's. While lacking Petrarch's polemical zeal and spirit of self-inquiry, he shared the Italian's intense love of language, his dislike for the complexities and pretenses of medieval institutions both secular and religious, and his commanding personal presence. More specifically, however, his ideas and overall direction betray the influence of Lorenzo Valla, whose works he treasured. Like Valla, who had attacked biblical textual criticism with a vengeance and proved the so-called Donation of Constantine to be a forgery, Erasmus contributed importantly to Christian philology. Also like Valla, he philosophically espoused a kind of Christian hedonism, justifying earthly pleasure from a religious perspective. But he was most like Valla (and indeed the entire rhetorical "arm" of Italian humanism) in giving philology prominence over philosophy. He described himself as a poet and orator rather than an inquirer after truth. His one major philosophical effort, a Christian defense of free will, was thunderously answered by Luther. Though his writings are a well of good sense, they are seldom profound and are predominantly derivative. In Latin eloquence, on the other hand, he was preeminent, both as stylist and theorist. His graceful and abundant Ciceronian prose (whose principles he set down in *De copia verborum et rerum*) helped shape the character of European style. Perhaps his most original work is *Moriae encomium* (*The Praise of*

The Praise of Folly

Folly), an elegant combination of satire and poetic insight whose influence was soon apparent in the work of More (to whom it was dedicated) and Rabelais.

The French humanists. Erasmus' associates in France included the influential humanists Robert Gaguin (1433–1501), Jacques Lefèvre d'Étaples (*c.* 1455–1536), and Guillaume Budé (Guglielmus Budaeus, 1467–1540). Of these three, Budé was most central to the development of French humanism, not only in his historical and philological studies but also in his use of his national influence to establish the Collège de France and the library at Fontainebleau. The influence of Francis I (1494–1547) and his learned sister Margaret of Angoulême (1492–1549) was important in fostering the new learning. The diversity and energy of French humanism is apparent in the activities of the Estienne family of publishers; the poetry of Pierre de Ronsard (1524–85), Joachim du Bellay (*c.* 1522–60), and Guillaume du Bartas (1544–90); the political philosophy of Jean Bodin (1530–96); the philosophical methodology of Petrus Ramus (Pierre de la Ramée, 1515–72); and the dynamic relationship between humanistic scholarship and church reform (see below, *Humanism and Christianity*). Hampered by religious repression and compressed more severely in time, the French movement lacked the intellectual fecundity and the programmatic unity of its Italian counterpart. In François Rabelais and Michel de Montaigne, however, the development of humanistic methods and themes resulted in unique and memorable achievement.

François Rabelais (c. 1490–1533). Rabelais ranks with Boccaccio as a founding father of Western realism. As a satirist and stylist (in his hands French prose became a free, poetic form), he influenced writers as important as Jonathan Swift, Laurence Sterne, and James Joyce and may be seen as a major precursor of modernism. His five books concerning the deeds of the giant princes Gargantua and Pantagruel constitute a treasury of social criticism, an articulate statement of humanistic values, and a forceful, if often outrageous, manifesto of human rights. Rabelaisian satire took aim at every social institution and (especially in Book III) every intellectual discipline. Broadly learned and unflaggingly alert to jargon and sham, he repeatedly focused on dogmas that fetter creativity, institutional structures that reward hypocrisy, educational traditions that inspire laziness, and philosophical methodologies that obscure elemental reality. His heroes, Gargantua and his son and heir Pantagruel, are figures whose colossal size and appetites (Rabelais's etymology for *Pantagruel* is "all-thirsty") symbolize the nobility and omnivorous curiosity that typified the humanistic scheme. The multifarious educational program detailed in *Gargantua* is reminiscent of Vittorino, Alberti, and the Montefeltro court; and the utopian Abbey of Thélème, whose gate bears the motto "Do as you please," is a tribute to enlightened will and pleasure in the manner of Valla, Erasmus, and More. Characteristically overstated and never wholly free of irony, Rabelais's work is a far cry from the earnest moral and educational programs of the early humanists. Rather than rebuild society, he seeks to amuse, edify, and refine it. His qualified endorsement of human dignity is based on the healthy balance of mind and body, the sanctity of all true learning, and the authenticity of direct experience.

Gargantua

Michel de Montaigne (1533–92). Montaigne's famous *Essays* are not only a compendious restatement and reevaluation of humanistic motives but also a milestone in the humanistic project of self-inquiry that had been originally endorsed by Petrarch. Scholar, traveler, soldier, and statesman, Montaigne was, like Machiavelli, alert to both theory and practice; but while Machiavelli saw practice as forming the basis for sound theory, Montaigne perceived in human events a multiplicity so overwhelming as to deny theoretical analysis. Montaigne's use of typical humanistic modalities—interpretation of the classics, appeals to direct experience, exclusive emphasis on the human realm, and universal curiosity—led him, in other words, to the refutation of a typical humanistic premise: that knowledge of the intellectual arts could teach one a sovereign art of life. In an effort to make his inquiry more inclusive and unsparing, Montaigne made himself the subject of his book, demonstrating through hundreds of personal anecdotes and admissions the ineluctable diversity of a single human spirit. His essays, which seem to move freely from one subject or viewpoint to another, are often in fact carefully organized dialectical structures that draw the reader, through thesis and antithesis, stated subject and relevant association, toward a multidimensional understanding of morality and history. The final essay, grandly titled "Of Experience," counsels a mature acceptance of life in all its contradictions. Human dignity, he implies, is indeed possible, but it lies less in heroic achievement than in painfully won self-knowledge. In this sense Montaigne's attitude toward the humanistic tradition is generally similar to that suggested in the work of Castiglione and Rabelais. While effectively taking issue with a number of the more extreme humanistic contentions, he retained and indeed justified the basic attitudes that gave the movement its form.

"Of Experience"

The English humanists. English humanism flourished in two stages: the first a basically academic movement that had its roots in the 15th century and culminated in the work of Sir Thomas More, Sir Thomas Elyot, and Roger Ascham, the second a poetic revolution led by Sir Philip Sidney and William Shakespeare.

Though continental humanists had held court positions since the days of Humphrey of Gloucester, English humanism as a distinct phenomenon did not emerge until late in the 15th century. At Oxford William Grocyn (*c.* 1446–1519) and his student Thomas Linacre (*c.* 1460–1524) gave impetus to a tradition of classical studies that would permanently influence English culture. Grocyn and Linacre attended Politian's lectures at the Platonic Academy of Florence. Returning to Oxford, they became central figures in a group that included such younger scholars as John Colet (1466/67–1519) and William Lily (1468?–1522). The humanistic contributions of the Oxford group were philological and institutional rather than philosophical or literary. Grocyn lectured on Greek and theology; Linacre produced several works on Latin grammar and translated Galen into Latin. To Linacre is owed the foundation of the Royal College of Physicians; to Colet, the foundation of St. Paul's School, London. Colet collaborated with Lily (the first headmaster of St. Paul's) and Erasmus in writing the school's constitution, and together the three scholars produced a Latin grammar (known alternately as "Lily's Grammar" and the "Eton Grammar") that would be central to English education for decades to come.

In Sir Thomas More (1478–1535), Sir Thomas Elyot (*c.* 1490–1546), and Roger Ascham (1515–68), English humanism bore fruit in major literary achievement. Educated at Oxford (where he read Greek with Linacre), More was also influenced by Erasmus, who wrote *The Praise of Folly* (Latin *Moriae encomium*) at More's house and named the book punningly after his English friend. More's famous *Utopia*, a kind of companion piece to *The Praise of Folly*, is similarly satirical of traditional institutions (Book I) but offers, as an imaginary alternative, a model society based on reason and nature (Book II). Reminiscent of Erasmus and Valla, More's Utopians eschew the rigorous cultivation of virtue and enjoy moderate pleasures, believing that "Nature herself prescribes a life of joy (that is, pleasure)" and seeing no contradiction between earthly enjoyment and religious piety. Significantly indebted both to classical thought and European humanism, the *Utopia* is also humanistic in its implied thesis that politics begins and ends with humanity: that politics is based exclusively on human nature and aimed exclusively at human happiness. Sir Thomas Elyot chose a narrower subject but developed it in more detail. His great work, *The Book Named The Governor*, is a lengthy treatise on the virtues to be cultivated by statesmen. Born of the same tradition that produced *The Prince* and *The Courtier*, *The Governor* is typical of English humanism in its emphasis on the accommodation of both classical and Christian virtues within a single moral view. Elyot's other contributions to English humanism include philosophical dialogues, moral essays, translations of ancient and contemporary writers (including Isocrates and Pico), an important Latin-English dictionary, and a highly popular health manual. He served

Utopia

his country as ambassador to the court of Charles V. Finally, the humanistic educational program set up at the turn of the century was vigorously supported by Sir John Cheke (1514–57) and codified by his student Roger Ascham. Ascham's famous pedagogical manual, *The Schoolmaster,* offers not only a complete program of humanistic education but also an evocation of the ideals toward which that education was directed.

Elizabeth I

Ascham had been tutor to the young princess Elizabeth, whose personal education was a model of humanistic pedagogy and whose writings and patronage bespoke great love of learning. Elizabeth I's reign (1558–1603) saw the last concerted expression of humanistic ideas. Elizabethan humanism, which added a unique element to the history of the movement, was the product not of pedagogues and philologists but of poets and playwrights.

Sidney and Spenser. Sir Philip Sidney (1554–86) was, like Alberti and Federico da Montefeltro, a living pattern of the humanistic ideal. Splendidly educated in the Latin classics at Shrewsbury and Oxford, Sidney continued his studies under the direction of the prominent French scholar Hubert Languet and was tutored in science by the learned John Dee. His brief career as writer, statesman, and soldier was of such acknowledged brilliance as to make him, after his tragic death in battle, the subject of an Elizabethan heroic cult. Sidney's major works, *Astrophel and Stella,* the *Defence of Poesie,* and the two versions of the *Arcadia,* are medleys of humanistic themes. In the sonnet sequence *Astrophel and Stella,* he surpassed earlier imitators of Petrarch by emulating not only the Italian humanist's subject and style but also his philosophical bent and habit of self-scrutiny. The *Defence of Poesie,* composed (like Erasmus' *Praise of Folly*) in the form of a classical oration, reasserts the theory of poetry as moral doctrine that had been articulated by Petrarch and Boccaccio and revived by the Italian Aristotelians of the 16th century. The later or "new" *Arcadia* is an epic novel whose theoretical concerns include the dualities of contemplation and action, reason and passion, and theory and practice. In this ambitious and unfinished work, Sidney attempts a characteristically humanistic synthesis of classical philosophy, Christian doctrine, psychological realism, and practical politics. Seen as a whole, moreover, Sidney's life and work form a significant contribution to a debate that had been smoldering since the decline of political liberty in Florence in the 15th century. How, it was asked, could humanism be politically active or "civic" in a Europe that was almost exclusively monarchic in structure? Many humanists had counseled retirement from active life, while Castiglione had seen his learned courtier rather as an advisor than as a leader. Sidney and his friend Edmund Spenser (1552/53–1599) sought to resolve this dilemma by creating a form of chivalric humanism. The image (taken on personally by Sidney and elaborated upon by Spenser in *The Faerie Queene*) of the hero as questing knight suggests that the humanist, even if not empowered politically, can achieve a valid form of activism by refining, upholding, and representing the values of a just and noble court. Spenser's poetic development of this humanistic program was even more specific than Sidney's. In his famous letter to Raleigh, he asserts that his purpose in *The Faerie Queene* is "to fashion a gentleman or noble person in virtuous and gentle discipline" and describes a project (never to be completed) of presenting his idea of the Aristotelian virtues in twelve poetic books. As with Sidney, however, this moral didacticism is neither self-righteous nor pedantic. The prescriptive content of *The Faerie Queene* is qualified by a strong emphasis on moral autonomy and a mature sense of the ambiguity of experience.

Chapman, Jonson, and Shakespeare. The poetry and drama of Shakespeare's time were a concourse of themes, ancient and modern, continental and English. Prominent among these motives were the characteristic topics of humanism. George Chapman (1559?–1634), the translator of Homer, was a forthright exponent of the theory of poetry as moral wisdom, holding that it surpassed all other intellectual pursuits. Ben Jonson (1572–1637) described his own humanistic mission when he wrote that a good poet was able "to inform young men to all good disciplines, inflame grown men to all great virtues, keep old men in their best and supreme state, or, as they decline to childhood, recover them to their first strength" and that the poet was "the interpreter and arbiter of nature, a teacher of things divine no less than human, a master in manners." Jonson, who sought this moral goal both in his tragedies and in his comedies, paid tribute to the humanistic tradition in *Catiline,* a tragedy in which Cicero's civic eloquence is portrayed in heroic terms.

Shakespeare

Less overtly humanistic, though in fact more profoundly so, was William Shakespeare (1564–1616). Thoroughly versed (probably at his grammar school) in classical poetic and rhetorical practice, Shakespeare early in his career produced strikingly effective imitations of Ovid and Plautus (*Venus and Adonis* and *The Comedy of Errors,* respectively) and drew on Ovid and Livy for his poem *The Rape of Lucrece.* In *Julius Caesar, Antony and Cleopatra,* and *Coriolanus* he developed Plutarchan biography into drama that, though Elizabethan in structure, is sharply classical in tone. Shakespeare clearly did not accept all the precepts of English humanism at face value. He grappled repeatedly with the problem of reconciling Christian doctrine with effective political action, and for a while (*e.g.,* in *Henry V*) seemed inclined toward the Machiavellian alternative. In *Troilus and Cressida,* moreover, he broadly satirized Chapman's Homeric revival and, more generally, the humanistic habit of idolizing classical heroism. Finally, he eschewed the moralism, rationalism, and self-conscious erudition of the humanists and was lacking as well in their fraternalism and their theoretical bent. Yet on a deeper level he must be acknowledged the direct and natural heir of Petrarch, Boccaccio, Castiglione, and Montaigne. Like them he delighted more in presenting issues than in espousing systems and held critical awareness, as opposed to doctrinal rectitude, to be the highest possible good. His plays reflect an inquiry into human character entirely in accord with the humanistic emphasis on the dignity of the emotions, and indeed it may be said that his unprecedented use of language as a means of psychological revelation gave striking support to the humanistic contention that language was the heart of culture and the index of the soul. Similarly, Shakespeare's unparalleled realism may be seen as the ultimate embodiment, in poetic terms, of the intense concern for specificity—be it in description, measurement, or imitation—endorsed across the board by humanists from Boccaccio and Salutati on. Shakespearean drama is a treasury of the disputes that frustrated and delighted humanism, including (among many others) action versus contemplation, theory versus practice, *res* versus *verbum,* monarchy versus republic, human dignity versus human depravity, and individualism versus communality. In treating of these polarities, he generally proceeds in the manner of Castiglione and Montaigne, presenting structures of balanced contraries rather than syllogistic endorsements of one side or another. In so doing, he achieves a higher realism, transcending the mere imitation of experience and creating, in all its conflict and fertility, a mirror of mind itself. Since the achievement of such psychological and cultural self-awareness was the primary goal of humanistic inquiry, and since humanists agreed that poetry was an uncommonly effective medium for this achievement, Shakespeare must be acknowledged as a preeminent humanist.

One cannot leave Shakespeare and the phenomenon of English humanism without reference to a highly important aspect of his later drama. Throughout his career, Shakespeare had shown a keen interest in the concept of art, not only as a general idea but also with specific reference to his own identity as dramatist. In two of his final plays, *The Winter's Tale* and *The Tempest,* he developed this concept into dramatic and thematic structures that had strongly doctrinal implications. Major characters in both plays practice a moral artistry—a kind of *humanitas* compounded of awareness, experience, imagination, compassion, and craft—that enables them to beguile and dominate other characters and to achieve enduring justice. This special skill, which is cognate with Shakespeare's own dramatic art, suggests a hypothetical solution to many of

the dilemmas posed in his earlier work. It implies that problems unavailable to political or religious remedy may be solved by creative innovation and that the art by which things are known and expressed may constitute, in and of itself, a valid field of inquiry and an instrument for cultural renewal. In developing this idea of the sovereignty of art, Shakespeare made the final major contribution to a humanistic tradition that will be discussed in the two sections that follow.

HUMANISM AND THE VISUAL ARTS

Humanistic themes and techniques were woven deeply into the development of Italian Renaissance art; conversely, the general theme of "art" was prominent in humanistic discourse. The mutually enriching character of the two disciplines is evident in a variety of areas.

Realism. Humanists paid conscious tribute to realistic techniques in art that had developed independently of humanism. Giotto di Bondone (*c.* 1266–1337), the Florentine painter responsible for the movement away from the Byzantine style and toward ancient Roman technique, was praised by Vasari as "the pupil of Nature." Giotto's own contemporary Boccaccio said of him in the *Decameron* that

Giotto

> there was nothing in Nature—the mother and ruling force of all created things with her constant revolution of the heavens—that he could not paint with his stylus, pen, or brush or make so similar to its original in Nature that it did not appear to be the original rather than a reproduction. Many times, in fact, in observing things painted by this man, the visual sense of men would err, taking what was painted to be the very thing itself.

Boccaccio, himself a naturalist and a realist, here subtly adopts the painter's achievement as a justification for his own literary style. So Shakespeare, at the end of the Renaissance, praises Giulio Romano (and himself), "who, had he himself eternity and could put breath into his work, would beguile Nature of her custom, so perfectly he is her ape" (*The Winter's Tale*). It should be noted that neither Vasari, Boccaccio, nor Shakespeare endorses realistic style as a *summum bonum:* realism is rather the means for regaining touch with the sovereign creative principle of Nature.

Classicism. Like the humanists, Italian artists of the 15th century saw a profound correlation between classical forms and realistic technique. Classical sculpture and Roman painting were emulated because of their ability to simulate perceived phenomena, while, more abstractly, classical myth offered a unique model for the artistic idealization of human beauty. Alberti, himself a close friend of Donatello and Brunelleschi, codified this humanistic theory of art, using the fundamental principle of mathematics as a link between perceived reality and the ideal. He developed a classically based theory of proportionality between architectural and human form, believing that the ancients sought "to discover the laws by which Nature produced her works so as to transfer them to the works of architecture."

Anthropocentricity and individualism. Humanism and Italian art were similar in giving paramount attention to human experience, both in its everyday immediacy and in its positive or negative extremes. The religious themes that dominated Renaissance art (partly because of generous church patronage) were frequently developed into images of such human richness that, as one contemporary observer noted, the Christian message was submerged. The human-centredness of Renaissance art, moreover, was not just a generalized endorsement of earthly experience. Like the humanists, Italian artists stressed the autonomy and dignity of the individual. High Renaissance art boasted a style of portraiture that was at once humanely appreciative and unsparing of detail. Heroes of culture such as Federico da Montefeltro and Lorenzo de' Medici, neither of whom was a conventionally handsome man, were portrayed realistically, as though a compromise with strict imitation would be an affront to their dignity as individuals. Similarly, artists of the Italian Renaissance were, characteristically, unabashed individualists. The biographies of Giotto, Brunelleschi, Leonardo, and Michelangelo by Giorgio Vasari (1511–74) not only describe artists who were well aware of their unique positions in society and history but also attest to a cultural climate in which, for the first time, the role of art achieved heroic stature. The autobiographical writings of the humanist Alberti, the scientist Gerolamo Cardano (1501–76), and the artist Benvenuto Cellini (1500–71) further attest to the individualism developing both in letters and in the arts: and Montaigne dramatized the analogy between visual mimesis and autobiographical realism when he said, in the preface to his *Essays,* that given the freedom he would have painted himself "*tout entier, et tout nu*" ("totally complete, and totally nude").

Biography and autobiography

Art as philosophy. Italian Renaissance painting, especially in its secular forms, is alive with visually coded expressions of humanistic philosophy. Symbol, structure, posture, and even colour were used to convey silent messages about humanity and nature. Renaissance style was so articulate, and the Renaissance sense of the unity of experience so deeply ingrained, that even architectural structures could be eloquently philosophical. Two features of Federico's palace at Urbino exemplify the profound interrelationship between humanistic principle and Renaissance art. The first feature is architectural. On the ground floor of the palace two private chapels, of roughly the same dimensions, stand side by side. The chapel at the left is a place of Christian worship, while that at the right is dedicated to the pagan Muses. Directly above these chapels is a study, the walls of which are covered with representations (in intarsia) of assorted humanistic heroes: Homer, Plato, Aristotle, Cicero, Virgil, Seneca, Boethius, St. Augustine, Dante, Petrarch, Bessarion, and Federico's revered teacher Vittorino, among others. The message conveyed by the positioning of the three rooms is hard to ignore. Devotion to the opposed principles of Christianity

SCALA/Art Resource, NY



Federico da Montefeltro, duke of Urbino, with his son Guidobaldo, portrait probably by Pedro Berruguete, 15th century. In the National Gallery of the Marches, Urbino, Italy.

and earthly (pagan) beauty is rendered possible by a humanistic learning (represented by the study) so generous and appreciative as to comprehend both extremes.

The second feature is iconographic—a portrait of Federico and his son Guidobaldo (probably by Pedro Berruguete) that occupies a central position on the wall of the study. It depicts the Duke, his full coat of armour partly covered by a courtly robe, sitting and reading. The son stands beside his father's chair, gazing out of the picture toward the viewer's left. An abbot's mitre rests on a shelf in the upper left, while the Duke's helmet sits on the floor in the lower right. Here also a typically humanistic message is evident. The Duke's scholarly attitude and curious attire suggest his triple role as warrior, ruler, and humanist. The two main axes of the picture—the line between mitre and helmet and the line between father and son—converge at the book, symbolizing the central role of humanistic learning in reconciling the concerns of church and state and in conveying humanistic virtue from generation to generation. The boy's outward gaze implies the characteristic direction of humanistic learning: into the world of action. The scope and organic wholeness of Federico's humanistic iconography are so striking as to rival great expressions of religious faith. The private heart of his palace concealed, like a genetic code, the principle that had given shape to the edifice and informed the state.

HUMANISM, ART, AND SCIENCE

It is impossible to speak knowledgeably about Renaissance science without first understanding the Renaissance concept of art. The Latin *ars* (inflected as *artis*) was applied indiscriminately to the verbal disciplines, mathematics, music, and science (the "liberal arts"), as well as to painting, sculpture, and architecture; it also could refer to technological expertise, to magic, and to alchemy. Any discipline involving the cultivation of skill and excellence was de facto an art. To the Renaissance, moreover, all arts were "liberal" arts in their capacity to "free" their practitioners to function effectively in specific areas. The art of rhetoric empowered the rhetorician to convince; the art of perspective empowered the painter to create visual illusion; the art of physics empowered the scientist to predict the force and motion of objects. "Art," in effect, was no more or less than articulate power, the technical or intellectual analogy to the political power of the monarch and the divine power of the god. The historical importance of this equation cannot be overestimated. If one concept may be said to have integrated all the varied manifestations of Renaissance culture and given organic unity to the period, it was this definition of art as power. With this definition in mind, one may understand why Renaissance humanists and painters assigned themselves such self-consciously heroic roles: in their artistic ability to delight, to captivate, to convince, they saw themselves as enfranchised directors and remakers of culture. One may also understand why a humanist-artist-scientist like Alberti would have seen no real distinction between the various disciplines he practiced. As profoundly interconnected means of understanding nature and humanity, and as media for effective reform and renewal, these disciplines were all components of an encompassing art. A similar point may be made about Machiavelli, who wrote a book about the "art" of warfare and who used history and logic to develop an art of government, or about the brilliant polymath Paracelsus, who spent his whole career perfecting an art that would comprehend all matter and all spirit. With the equation of art and power in mind, finally, one may understand why a revolutionary scientist like Galileo (1564–1642) put classical and medieval science through a winnowing fan, keeping only such components as allowed for physically reproducible results. Since every Renaissance art aimed for a dominion or conquest, it was completely appropriate that science should leave its previously contemplative role and focus upon the conquest of nature.

The "liberal arts"

Humanism benefited the development of science in a number of more specific ways. Alberti's technological applications of mathematics, and his influential statement that mathematics was the key to all sciences, grew out of his humanistic education at Padua. Vittorino, another student at Padua, went on to make mathematics a central feature of his educational program. Gerolamo Cardano, a scholar of renowned humanistic skills, made major contributions to the development of algebra. In short, the importance of mathematics in humanistic pedagogy and the fact that major humanists like Vittorino and Alberti were also mathematicians may be seen as contributing to the critical role mathematics would play in the rise of modern science. Humanistic philology, moreover, supplied scientists with clean texts and clear Latin translations of the classical works—Plato, Aristotle, Euclid, Archimedes, and even Ptolemy—that furthered their studies. The richness of the classical heritage in science is often underestimated. Galileo, who considered Archimedes his mentor, also prized the dialogues of Plato, in particular the *Meno*. The German philosopher Ernst Cassirer has demonstrated the likelihood that Galileo was fond of the *Meno* because it contained the first statement of the "hypothetical" method, a modus operandi that characterized Galileo's own scientific practice and that would come to be known as one of the chief principles of the New Science. Humanism may also be seen as offering, of itself, methods and attitudes suitable for application in nonhumanistic fields. It might be argued, for example, that the revolutionary social science of Machiavelli and Juan Luis Vives (1492–1540) was due in large measure to their application of humanistic techniques to fields that lay outside the normal purview of humanism. But most of all it was the general spirit of humanism—critical, questing, ebullient, precise, focused on the physical world, and passionate in its quest for results—that fostered the development of the scientific spirit in social studies and natural philosophy.

The classical heritage in science

HUMANISM AND CHRISTIANITY

Though much humanistic activity was specifically Christian in intention, and though the majority of humanists made firm avowals of faith, the relationship between Christianity and humanism is complex and not wholly untroubled. First, humanists from Petrarch onward recognized that the classical (pagan) direction of humanism necessarily constituted, if not a challenge to Christianity, at least a breach in the previous totality of Christian devotion. The Christian truth that had been acknowledged as comprehending all phenomena, earthly or heavenly, now had to coexist with a classical attitude that was overwhelmingly directed toward earthly life. Humanistic efforts to resolve the contradictions implied by these two attitudes were, if one may judge by their variety, never wholly successful. In particular, the extent to which humanistic inquiry led scholars toward the secular realm, and the extent to which humanistic pedagogy concentrated on secular subjects, suggest erosions of the domain of faith. Coluccio Salutati, who urged the young Poggio not to let humanistic enthusiasm take precedence over Christian piety, thereby acknowledged a dualism implicit in the humanistic program and never wholly absent from its historical development.

Implicit dualism

Second, the humanistic philology that meticulously compared ancient sources and "cleaned up" the texts of important Christian writings was a serious challenge to the authority of the church. With new authorities or refined texts in hand, humanists found fault with established commentaries and questioned traditional interpretations. Valla's arraignment of the Donation of Constantine and Bessarion's discovery that the supposed Dionysius the Areopagite (later called Pseudo-Dionysius) had borrowed some of his material from Plato exemplify the uneasy relationship between humanism and Catholic dogma. Third, the independent and broadly critical attitude innate to humanism could not but threaten the unanimity of Christian belief. Intellectual individualism, which has never been popular in any church, put particular stress on a religion that encouraged simple faith and alleged universal authority. Finally, humanism repeatedly fostered the impulse of religious reform. The humanistic emphasis on total authenticity and direct contact with sources had, as its religious correlative, a desire to obliterate the medieval accretions and procedural complexities that stood between the worshiper and his god. The reform-mindedness of

The impulse of reform

such humanists as Petrarch, Boccaccio, Erasmus, and Rabelais was balanced on the religious side by reformers such as Calvin and Melanchthon, who employed humanistic techniques in their own cause. And the reform movement, while it may have modernized and thus preserved Christianity, rang the death knell for a medieval culture whose essential characteristic had been participation in a universal church.

LATER FORTUNES OF HUMANISM

Shakespeare may be seen as the last major interpreter of the humanistic program. Sir Francis Bacon and John Milton, though formidably adept at humanistic techniques, diverged in their major work from the central current of humanism, Bacon toward natural science, Milton toward theology. If Bacon's rationalism may be seen as a link between humanism and the Enlightenment, his strong emphasis on nature (rather than humanity) as subject matter presaged the permanent separation of the sciences from the humanities. In Milton's theocentricity, on the other hand, lay the Christian distrust (going back, perhaps, to Luther) of humanistic secularism. These epochal divergences, moreover, were complemented by a series of rifts and ramifications within the humanistic movement. The split between philosophy and letters was, over future generations, to be compounded by the development of countless discrete specialties within both fields. Philosophers came more and more to define themselves within narrow boundaries. Creative writers and "critics" took up distinct positions and assumed adversarial relationships. The profound loss of coherence in humane letters was furthered by the gradual decline of Latin as the lingua franca of European intellectuals and the consequent separation of national traditions.

By the 19th century, humanism was such a lost art as to have to be reassembled, like a disjointed fossil, by careful historians. Of course there were exceptions. Jonathan Swift (1667–1745) reasserted humanistic values in a broad-based attack on contemporary institutions, and in Gottfried Wilhelm Leibniz (1646–1716) can be found the serious intention and multifarious curiosity that characterized humanism at its best. Strong humanistic motives may be found in Germany at the turn of the 19th century, particularly in the work of Gotthold Ephraim Lessing (1729–81), Friedrich von Schiller (1759–1805), and Georg Wilhelm Friedrich Hegel (1770–1831); while Johann Wolfgang von Goethe (1749–1832) was perhaps the last individual whose breadth of achievement and sense of the unity of experience lived up to the ideal established by Alberti.

Goethe

More recently, the mode of inquiry and interpretation developed by the political philosopher Leo Strauss (1899–1973) showed strong signs of the humanistic spirit. But in general the traces of the original program have been scattered. To the modern mind, a "humanist" is a university scholar, walled off from the interdisciplinary scope of the original humanistic program and immune to the active experience that was its basis and its goal. This decline is easy enough to explain. Had there been nothing else, one external factor would have made the cultivation of *humanitas*, as originally practiced, more and more difficult from the beginning of the 16th century on. The proliferation of published work in all fields, and the creation of many new fields, made increasingly impracticable the development of the comprehensive learning and awareness that were central to the original program. In 1500 the major texts constituting a humanistic education, though numerous, could still be counted; by 1900 they were legion, and people had long ceased agreeing about exactly which ones they were. But problems implicit in the movement were equally responsible for its demise. The characteristic emphases on rhetoric and philology, which gave the humanistic movement vitality and made it available to countless students of moderate gifts, also betokened its impermanence. Weak in dialectic or any other comprehensively analytic method, the movement had no instrument for self-examination, no medium for self-renewal. By the same token, neither had humanism any valid means of defense against the attackers—scientists, fundamentalists, materialists, and others—who camped in ever larger numbers on its borders. Lacking an integral method, finally, humanism in effect lacked a centre and became prey to an endless series of ramifications. While eloquent humanists rambled through Europe and spread the word about the classics, the method that might have unified their efforts lay, available but unheeded, in texts of Plato and Aristotle. Given this core of rigorous analysis, humanism might (all other challenges notwithstanding) have retained its basic character for centuries. But ironically it might also have failed to attract followers.

The lack of method

CONCLUSION

Though lacking permanence itself, humanism in large measure established the climate and provided the medium for the rise of modern thought. An impressive variety of major developments in literature, philosophy, art, religion, social science, and even natural science had their basis in humanism or were significantly nourished by it. Important spokesmen in all fields regularly made use of humanistic eloquence to further their causes. More generally, the so-called modern awareness—that sense of alienation and freedom applied both to the individual and to the race—derives ultimately, for better or worse, from humanistic sources. But with humanism, as with every other historical subject, one should beware lest valid concern about changes, crises, sources, and influences obscure the even more important issues of human continuity and human value. Whatever its weaknesses and inner conflicts, the humanistic movement was heroic in its breadth and energy, remarkable in its aspirations. For human development in all fields, it created a context of seldom-equaled fertility. Its characteristic modalities of thought, speech, and image lent themselves to the promptings of genius and became the media for enduring achievement. Its moral program formed the basis for lives that are remembered with admiration.

BIBLIOGRAPHY

General treatments: The three general studies most helpful in approaching the phenomenon of humanism are EUGENIO GARIN, *Italian Humanism: Philosophy and Civic Life in the Renaissance*, trans. from the Italian (1965, reprinted 1975); PAUL OSKAR KRISTELLER, *Renaissance Thought and Its Sources* (1979); and CHARLES TRINKAUS, *The Scope of Renaissance Humanism* (1983). Garin's book is probably the most unified and incisive treatment of Italian humanism yet produced, while Kristeller and Trinkaus offer extremely well-documented analyses of major issues in the history and historiography of humanism. Other valuable readings include HANS BARON, *The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in an Age of Classicism and Tyranny*, rev. ed. (1966); QUIRINUS BREEN, *Christianity and Humanism: Studies in the History of Ideas* (1968); JACOB BURCKHARDT, *The Civilization of the Period of the Renaissance in Italy: An Essay* (1878; originally published in German, 1860), available in later English-language editions; DOUGLAS BUSH, *The Renaissance and English Humanism* (1939, reprinted 1972); ERNST CASSIRER, *The Individual and the Cosmos in Renaissance Philosophy* (1964, reprinted 1972; originally published in German with appendices, 1927); JACK D'AMICO, *Knowledge and Power in the Renaissance* (1977); MYRON P. GILMORE, *The World of Humanism, 1453–1517* (1952; reprinted 1983); DENYS HAY, *The Italian Renaissance in Its Historical Background*, 2nd ed. (1977); PAUL OSKAR KRISTELLER, *Renaissance Concepts of Man, and Other Essays* (1972); EDWARD P. MAHONEY (ed.), *Philosophy and Humanism* (1976); ROBERT MANDROU, *From Humanism to Science, 1480 to 1700* (1979; originally published in French, 1973); HEIKO A. OBERMAN and THOMAS A. BRADY, JR. (eds.), *Itinerarium Italicum: The Profile of the Italian Renaissance in the Mirror of Its European Transformations* (1975); CHARLES B. SCHMITT, *Studies in Renaissance Philosophy and Science* (1981); JOHN ADDINGTON SYMONDS, *Renaissance in Italy*, 7 vol. (1875–86, reprinted 1971–72); GUISSEPPE TOFFANIN, *History of Humanism* (1954; originally published in Italian, 1933); BERTHOLD L. ULLMAN, *Studies in the Italian Renaissance*, 2nd ed. (1975); and ROBERTO WEISS, *The Spread of Italian Humanism* (1964). For classical and medieval backgrounds, see ERNST ROBERT CURTIUS, *European Literature and the Latin Middle Ages* (1973; originally published in German, 1948); MOSES HADAS, *Humanism: The Greek Ideal and Its Survival* (1960, reprinted 1972); and WERNER JAEGER, *Paideia: The Ideals of Greek Culture*, 2nd ed., 3 vol. (1965; originally published in German, 1934).

Specific topics: T.W. BALDWIN, *William Shakspere's Small*

Latine and Lesse Greeke (1944, reprinted 1966); HANS BARON, *From Petrarch to Leonardo Bruni: Studies in Humanistic and Political Literature* (1968); GENE BRUCKER, *Renaissance Florence* (1969, reprinted 1983); ERNST CASSIRER, "Galileo's Platonism," in M.F. ASHLEY MONTAGU (ed.), *Studies and Essays in the History of Science and Learning*, pp. 277–297 (1946; reprinted 1975); JOAN GADOL, *Leon Battista Alberti* (1969); EUGENIO GARIN, *Science and Civic Life in the Italian Renaissance* (1969, reissued 1978; originally published in Italian, 1965; 4th Italian ed., 1980); PAUL OSKAR KRISTELLER and PHILIP P. WIENER (eds.), *Renaissance Essays* (1968); LAURO MARTINES, *The Social World of the Florentine Humanists, 1390–1460* (1963); JAMES J. MURPHY (ed.), *Renaissance Eloquence: Studies in the Theory and Practice of Renaissance Rhetoric* (1983); IRWIN PANOFSKY, *Renaissance and Renascences in Western Art*, 2 vol. (1960, reissued in 1 vol, 1969); J.H. PLUMB (ed.), *Renaissance Profiles* (1965); PASQUALE ROTONDI, *The Ducal Palace of Urbino: Its Architecture and Decoration* (1969; originally published in Italian in 2 vol., 1950–51); CHARLES B. SCHMITT, *Aristotle and the Renaissance* (1983); CHARLES TRINKAUS, *In Our Image and Likeness: Humanity and Divinity in Italian Humanist Thought*, 2 vol. (1970), and *The Poet as Philosopher: Petrarch and the Formation of Renaissance Consciousness* (1979); BERTHOLD L. ULLMAN, *The Humanism of Coluccio Salutati* (1963); WILLIAM HARRISON WOODWARD, *Vittorino da Feltre and Other Humanist Educators* (1897, reissued 1970), and *Studies in Education During the Age of the Renaissance, 1400–1600* (1906, reissued 1967); and G.F. YOUNG, *The Medici*, 2 vol. (1909, reissued in 1 vol., 1933).

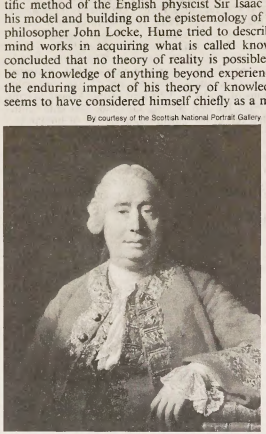
Works of the humanists: Works by the later humanists (*c.* 1500 and after) and the English poet-humanists mentioned in the article, including Castiglione, Cellini, Elyot, Erasmus, Jonson, Machiavelli, Montaigne, More, Pico della Mirandola, Rabelais, Shakespeare, Sidney, Spenser, Tasso, and Vasari, are readily available in many modern English editions. For the writings of the earlier humanists, see LEON BATTISTA ALBERTI, *The Family in Renaissance Florence*, trans. by RENÉE NEU WATKINS (1969); GIOVANNI BOCCACCIO, *The Decameron*, trans. by MARK MUSA and PETER BONDANELLA (1982), and *Boccaccio on Poetry*, 2nd ed., ed. and trans. by CHARLES G. OSGOOD (1956, reprinted 1978), a translation of the preface and books 14 and 15 of his *De genealogia deorum gentilium;* biographies of Dante by Boccaccio and Leonardo Bruni in *The Earliest Lives of Dante*, trans. by JAMES ROBINSON SMITH (1901, reprinted 1976); letters by Poggio Bracciolini in *Two Renaissance Book Hunters*, trans. by PHYLLIS WALTER GOODHART GORDAN (1974); and works by Petrarch, including *The Life of Solitude*, trans. by JACOB ZEITLIN (1924, reprinted 1978); "On His Own Ignorance and That of Many Others," trans. by HANS NACHOD in ERNST CASSIRER, PAUL OSKAR KRISTELLER, and JOHN HERMAN RANDALL, JR. (eds.), *The Renaissance Philosophy of Man* (1948, reprinted 1971); and *Petrarch's Secret; or, The Soul's Conflict with Passion*, trans. by WILLIAM H. DRAPER (1911, reprinted 1978). WILLIAM HARRISON WOODWARD, *op. cit.*, contains valuable translations of works by Vergerio, Bruni, Aeneas Sylvius Piccolomini (Pius II), and Battista Guarino. Sixteenth-century writings that reflect the breadth and vitality of humanistic attitudes are JUAN LUIS VIVES, *On Education*, trans. by FOSTER WATSON (1913, reprinted 1971); GIROLAMO CARDANO, *The Book of My Life*, trans. by JEAN STONER (1930, reprinted 1962); TORQUATO TASSO, *Tasso's Dialogues*, trans. by CARNES LORD and DAIN A. TRAFTON (1982); and PARACELSUS, *Selected Writings*, trans. by NORBERT GUTERMAN, 2nd rev. ed. (1958; reissued with a new bibliography, 1969; originally published in German, 1942, ed. by JOLANDE JACOBI).

(Ro.G.)

Hume

David Hume was an 18th-century Scottish empiricist philosopher, historian, economist, and essayist, who conceived of philosophy as the inductive, experimental science of human nature. Taking the scientific method of the English physicist Sir Isaac Newton as his model and building on the epistemology of the English philosopher John Locke, Hume tried to describe how the mind works in acquiring what is called knowledge. He concluded that no theory of reality is possible; there can be no knowledge of anything beyond experience. Despite the enduring impact of his theory of knowledge, Hume seems to have considered himself chiefly as a moralist.

By courtesy of the Scottish National Portrait Gallery



Hume, oil painting by Allan Ramsay, 1766. In the Scottish National Portrait Gallery, Edinburgh.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Ten, Division V, especially Section 51.

Early life and works. Hume was the younger son of Joseph Hume, the modestly circumstanced laird, or lord, of Ninewells, a small estate adjoining the village of Chirnside, about nine miles distant from Berwick-upon-Tweed on the Scottish side of the border. David's mother, Catherine, a daughter of Sir David Falconer, president of the Scottish court of session, was in Edinburgh when he was born, on May 7 (April 26, old style), 1711. In his third year his father died. He entered Edinburgh University when he was about 12 years old and left it at 14 or 15, as was then usual. Pressed a little later to study law (in the family tradition on both sides), he found it distasteful and instead read voraciously in the wider sphere of letters. Because of the intensity and excitement of his intellectual discovery, he had a nervous breakdown in 1729, from which it took him a few years to recover.

In 1734, after trying his hand in a merchant's office in Bristol, he came to the turning point of his life and retired to France for three years. Most of this time he spent at La Flèche on the Loire, in the old Anjou, studying and writing *A Treatise of Human Nature.* The *Treatise* was Hume's attempt to formulate a full-fledged philosophical system. It is divided into three books: book I, on understanding, aims at explaining man's process of knowing, describing in order the origin of ideas, the ideas of space and time, causality, and the testimony of the senses; book II, on the "passions" of man, gives an elaborate psychological machinery to explain the affective, or emotional, order in man and assigns a subordinate role to reason in this mechanism; book III, on morals, describes moral goodness in terms of "feelings" of approval or disapproval that a person has when he considers human behaviour in the light of the agreeable or disagreeable consequences either to himself or to others. Although the *Treatise* is Hume's most thorough exposition of his thought, at the end of his life he vehemently repudiated it as juvenile, avowing that only his later writings presented his considered views. The *Treatise* is not well constructed, in parts oversubtle, confusing because of ambiguity in important terms (especially "reason"), and marred by willful extravagance of statement and rather theatrical personal avowals. For these reasons his mature condemnation of it was perhaps not entirely misplaced. Book I, nevertheless, has been more read in academic circles than any other of his writings.

Writing the *Treatise*

Returning to England in 1737, he set about publishing the *Treatise.* Books I and II were published in two volumes in 1739; book III appeared the following year. The poor reception of this, his first and very ambitious work, depressed him; but his next venture, *Essays, Moral and Political* (1741–42), won some success. Perhaps encouraged by this, he became a candidate for the chair of moral philosophy at Edinburgh in 1744. Objectors alleged heresy and even atheism, pointing to the *Treatise* as evidence. Unsuccessful, Hume left the city, where he had been living since 1740, and began a period of wandering: a sorry year near St. Albans as tutor to the mad marquess of Annandale (1745–46); a few months as secretary to Gen. James St. Clair (a member of a prominent Scottish family), with whom he saw military action during an abortive expedition to Brittany (1746); a little tarrying in London and at Ninewells; and then some further months with General St. Clair on an embassy to the courts of Vienna and Turin (1748–49).

Mature works. During his years of wandering Hume was earning the money that he needed to gain leisure for his studies. Some fruits of these studies had already appeared before the end of his travels, viz., a further *Three Essays, Moral and Political* (1748) and *Philosophical Essays Concerning Human Understanding* (1748). The latter is a rewriting of book I of the *Treatise* (with the addition of his essay "On Miracles," which became notorious for its denial that a miracle can be proved by any amount or kind of evidence); it is better known as *An Enquiry Concerning Human Understanding,* the title Hume gave to it in a revision of 1758. The *Enquiry Concerning the Principles of Morals* (1751) was a rewriting of book III of the *Treatise.* It was in these works that Hume expressed his mature thought.

Theory of knowledge

An Enquiry Concerning Human Understanding is an attempt to define the principles of human knowledge. It poses in logical form significant questions about the nature of reasoning in regard to matters of fact and experience, and it answers them by recourse to the principle of association. The basis of his exposition is a twofold classification of objects of awareness. In the first place, all such objects are either "impressions," data of sensation or of internal consciousness, or "ideas," derived from such data by compounding, transposing, augmenting, or diminishing. That is to say, the mind does not create any ideas but derives them from impressions. From this Hume develops a theory of meaning. A word that does not stand directly for an impression has meaning only if it brings before the mind an object that can be gathered from an impression by one of the mental processes mentioned. In the second place, there are two approaches to construing meaning, an analytical one, which concentrates on the "relations of ideas," and an empirical one, which focuses on "matters of fact." Ideas can be held before the mind simply as meanings, and their logical relations to one another can then be detected by rational inspection. The idea of a plane triangle, for example, entails the equality of its internal

angles to two right angles, and the idea of motion entails the ideas of space and time, irrespective of whether there really are such things as triangles and motion. Only on this level of mere meanings, Hume asserts, is there room for demonstrative knowledge. Matters of fact, on the other hand, come before the mind merely as they are, revealing no logical relations; their properties and connections must be accepted as they are given. That primroses are yellow, that lead is heavy, and that fire burns things are facts, each shut up in itself, logically barren. Each, so far as reason is concerned, could be different: the contradictory of every matter of fact is conceivable. Therefore, any demonstrative science of fact is impossible.

Hume's doctrine of causality and belief

From this basis Hume develops his doctrine about causality. The idea of causality is alleged to assert a necessary connection among matters of fact. From what impression, then, is it derived? Hume states that no causal relation among the data of the senses can be observed, for, when a person regards any events as causally connected, all that he does and can observe is that they frequently and uniformly go together. In this sort of togetherness it is a fact that the impression or idea of the one event brings with it the idea of the other. A habitual association is set up in the mind; and, as in other forms of habit, so in this one, the working of the association is felt as compulsion. This feeling, Hume concludes, is the only discoverable impressional source of the idea of causality.

Hume then considers the process of causal inference, and in so doing he introduces the concept of belief. When a person sees a glass fall, he not only thinks of its breaking but expects and believes that it will break; or, starting from an effect, when he sees the ground to be generally wet, he not only thinks of rain but believes that there has been rain. Thus belief is a significant component in the process of causal inference. Hume then proceeds to investigate the nature of belief, claiming that he was the first to do so. He uses this term in the narrow sense of belief regarding matters of fact. He defines belief as a sort of liveliness or vividness that accompanies the perception of an idea. A belief is more than an idea; it is a vivid or lively idea. This vividness is originally possessed by some of the objects of awareness, by impressions and the simple memory images of them. By association it comes to belong to certain ideas as well. In the process of causal inference, then, an observer passes from an impression to an idea regularly associated with it. In the process the aspect of liveliness proper to the impression infects the idea, Hume asserts. And it is this aspect of liveliness that Hume defines as the essence of belief.

Hume does not claim to prove that the propositions, (1) that events themselves are causally related and (2) that they will be related in the future in the same ways as they were in the past, are false. He firmly believed both of these propositions and insisted that everybody else believed them, will continue to believe them, and must continue to believe them in order to survive. They are natural beliefs, inextinguishable propensities of human nature, madness apart. What Hume claims to prove is that natural beliefs are not obtained and cannot be demonstrated either by empirical observation or by reason, whether intuitive or inferential. Reflection shows that there is no evidence for them and shows also both that we are bound to believe them and that it is sensible or sane to do so.. This is Hume's skepticism: it is an affirmation of that tension, a denial not of belief but of certainty.

Theory of morals and historical writing

The *Enquiry Concerning the Principles of Morals* is a refinement of Hume's thinking on morality, in which he views sympathy as the fact of human nature lying at the basis of all social life and personal happiness. Defining morality as those qualities that are approved (1) in whomsoever they happen to be and (2) by virtually everybody, he sets himself to discover the broadest grounds of the approvals. He finds them, as he found the grounds of belief, in "feelings," not in "knowings." Moral decisions are grounded in moral sentiment. Qualities are valued either for their utility or for their agreeableness, in each case either to their owners or to others. Hume's moral system aims at the happiness of others (without any such formula as "the greatest happiness of the greatest number") and at the happiness of self. But regard for others accounts for the greater part of morality. His emphasis is on altruism: the moral sentiments that he claims to find in human beings, he traces, for the most part, to a sentiment for and a sympathy with one's fellows. It is human nature, he holds, to laugh with the laughing and to grieve with the grieved and to seek the good of others as well as one's own. Two years after the *Enquiry* was published, Hume confessed, "I have a partiality for that work"; and at the end of his life he judged it "of all my writings incomparably the best." Such statements, along with other indications in his later writings, make it possible to suspect that he regarded his moral doctrine as his major work. He here writes as a man having the same commitment to duty as his fellows. The traditional view that he was a detached scoffer is deeply wrong; he was skeptical not of morality but of much theorizing about it.

Following the publication of these works, Hume spent several years (1751–63) in Edinburgh, with two breaks in London. An attempt was made to get him appointed as successor to Adam Smith, the Scottish economist (later to be his close friend), in the chair of logic at Glasgow, but the rumour of atheism prevailed again. In 1752, however, Hume was made keeper of the Advocates' Library at Edinburgh. There, "master of 30,000 volumes," he could indulge a desire of some years to turn to historical writing. His *History of England,* extending from Caesar's invasion to 1688, came out in six quarto volumes between 1754 and 1762, preceded by *Political Discourses* (1752). His recent writings had begun to make him known, but these two brought him fame, abroad as well as at home. He also wrote *Four Dissertations* (1757), which he regarded as a trifle, although it included a rewriting of book II of the *Treatise* (completing his purged restatement of this work) and a brilliant study of "the natural history of religion." In 1762 James Boswell, the biographer of Samuel Johnson, called Hume "the greatest writer in Britain," and the Roman Catholic Church, in 1761, paid him the attention of putting all his writings on the *Index,* its list of forbidden books.

Hume's reception in Paris

The most colourful episode of his life ensued: in 1763 he left England to become secretary to the British embassy in Paris under the Earl of Hertford. The society of Paris accepted him, despite his ungainly figure and gauche manner. He was honoured as eminent in breadth of learning, in acuteness of thought, and in elegance of pen and was taken to heart for his simple goodness and cheerfulness. The salons threw open their doors to him, and he was warmly welcomed by all. For four months in 1765 he acted as chargé d'affaires at the embassy. When he returned to London at the beginning of 1766 (to become, a year later, undersecretary of state), he brought Jean-Jacques Rousseau, the Swiss philosopher connected with the *Encyclopédie* of Diderot and d'Alembert, with him and found him a refuge from persecution in a country house at Wootton in Staffordshire. This tormented genius suspected a plot, took secret flight back to France, and spread a report of Hume's bad faith. Hume was partly stung and partly persuaded into publishing the relevant correspondence between them with a connecting narrative (*A Concise and Genuine Account of the Dispute Between Mr. Hume and Mr. Rousseau,* 1766).

In 1769, somewhat tired of public life and of England too, he again established a residence in his beloved Edinburgh, deeply enjoying the company—at once intellectual and convivial—of friends old and new (he never married), as well as revising the text of his writings. He issued five further editions of his *History* between 1762 and 1773 as well as eight editions of his collected writings (omitting the *Treatise, History,* and ephemera) under the title *Essays and Treatises* between 1753 and 1772, besides preparing the final edition of this collection, which appeared posthumously (1777), and *Dialogues Concerning Natural Religion,* held back under pressure from friends and not published until 1779. His curiously detached autobiography, *The Life of David Hume, Esquire, Written by Himself* (1777; the title is his own), is dated April 18, 1776. After a long illness he died in his Edinburgh house on Aug. 25, 1776, and was buried on Calton Hill.

Adam Smith, his literary executor, added to the *Life* a letter that concludes with his judgment on his friend as "approaching as nearly to the idea of a perfectly wise and virtuous man as perhaps the nature of human frailty will permit." His distinguished friends, with ministers of religion among them, certainly admired and loved him, and there were younger men indebted either to his influence or to his pocket. The mob had heard only that he was an atheist and simply wondered how such an ogre would manage his dying. Yet Boswell has recounted, in a passage in his *Private Papers,* that, when he visited Hume in his last illness, the philosopher put up a lively, cheerful defense of his disbelief in immortality.

Significance and influence. That Hume was one of the major figures of his century can hardly be doubted. So his contemporaries thought, and his achievement, as seen in historical perspective, confirms that judgment, though with a shift of emphasis. Some of the reasons for the assessment may be given under four heads:

As a writer. Hume's style was praised in his lifetime and has often been praised since. It exemplifies the classical standards of his day. It lacks individuality and colour, for he was always proudly on guard against his emotions. The touch is light, except on slight subjects, where it is rather heavy. Yet in his philosophical works he gives an unsought pleasure. Here his detachment, levelness (all on one plane), smoothness, and daylight clearness are proper merits. It is as one of the best writers of scientific prose in English that he stands in the history of style.

As a historian. Library catalogs still list Hume as "Hume, David, the Historian." Between his death and 1894, there were at least 50 editions of his *History;* and an abridgment, *The Student's Hume* (1859; often reprinted), remained in common use for 50 years. Though now outdated, Hume's *History* must be regarded as an event of cultural importance. In its own day, moreover, it was an innovation, soaring high above its very few predecessors. It was fuller and set a higher standard of impartiality. His *History of England* not only traced the deeds of kings and statesmen but also displayed the intellectual interests of the educated citizens, as may be seen, for instance, in the pages on literature and science under the Commonwealth at the end of chapter 3 and under James II at the end of chapter 2. It was unprecedentedly readable, in structure as well as in phrasing. Persons and events were woven into causal patterns that furnished a narrative with the goals and resting points of recurrent climaxes. That was to be the plan of future history books for the general reader.

As an economist. Hume steps forward as an economist in the *Political Discourses* incorporated in *Essays and Treatises* as part 2 of *Essays Moral and Political.* How far he influenced his friend Adam Smith, 12 years his junior, remains uncertain: they had broadly similar principles, and both had the excellent habit of illustrating and supporting these from history. He did not formulate a complete system of economic theory, as did Adam Smith in his *Wealth of Nations,* but Hume introduced several of the new ideas around which the "classical economics" of the 18th century was built. His level of insight can be gathered from his main contentions: that wealth consists not of money but of commodities; that the amount of money in circulation should be kept related to the amount of goods in the market (two points made by Berkeley); that a low rate of interest is a symptom not of superabundance of money but of booming trade; that no nation can go on exporting only for bullion; that each nation has special advantages of raw materials, climate, and skill, so that a free interchange of products (with some exceptions) is mutually beneficial; and that poor nations impoverish the rest just because they do not produce enough to be able to take much part in that exchange. He welcomed advance beyond an agricultural to an industrial economy as a precondition of any but the barer forms of civilization.

Main contentions in economics

As a philosopher. Hume conceived of philosophy as the inductive science of human nature, and he concluded that man is more a creature of sensitive and practical sentiment than of reason. On the Continent he is seen as one of the few British classical philosophers. For some Germans his importance lies in the fact that Immanuel Kant conceived his critical philosophy in direct reaction to Hume. Hume was one of the influences that led Auguste Comte, the 19th-century French mathematician and sociologist, to positivism. In Britain his positive influence is seen in Jeremy Bentham, the early 19th-century jurist and philosopher, who was moved to utilitarianism (the moral theory that right conduct should be determined by the usefulness of its consequences) by book III of the *Treatise,* and more extensively in John Stuart Mill, the philosopher and economist who lived later in the 19th century.

In throwing doubt on the assumption of a necessary link between cause and effect, Hume was the first philosopher of the postmedieval world to reformulate the skepticism of the ancients. His reformulation, moreover, was carried out in a new and compelling way. Although Hume admired Newton, Hume's subtle undermining of causality called in question the philosophical basis of Newton's science as a way of looking at the world, inasmuch as this rested on the identification of a few fundamental causal laws that govern the universe. As a result the positivists of the 19th century were obliged to wrestle with Hume's questioning of causality if they were to succeed in their aim of making science the central framework of human thought. In the 20th century it was Hume's naturalism rather than his skepticism that attracted attention, chiefly among analytic philosophers. Hume's naturalism lies in his belief that philosophical justification could only be rooted in regularities of the natural world. The attraction of this for analytic philosophers was that it seemed to provide a solution to the problems arising from the skeptical tradition that Hume himself, in his other philosophical role, had done so much to reinvigorate.

MAJOR WORKS

PHILOSOPHY AND RELIGION: *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects* (1739–40); *An Abstract of a Book Lately Published: Entitled, A Treatise of Human Nature, etc., Wherein the Chief Argument of That Book Is Farther Illustrated and Explained* (1740); *Philosophical Essays Concerning Human Understanding* (1748; many later editions entitled *An Enquiry Concerning Human Understanding*); *Four Dissertations* (1757); *Dialogues Concerning Natural Religion* (1779); *A Letter from a Gentleman to His Friend Containing Some Observations on Religion and Morality* (1745).

POLITICS AND MORALS: *Essays, Moral and Political* (1741–42); *An Enquiry Concerning the Principles of Morals* (1751); *Political Discourses* (1752).

HISTORY: *The History of Great Britain* (1754–57); *The History of England Under the House of Tudor* (1759); *The History of England from the Invasion of Julius Caesar to the Accession of Henry VII* (1762).

OTHER WORKS: *A Concise and Genuine Account of the Dispute Between Mr. Hume and Mr. Rousseau* (1766); *The Life of David Hume, Esquire, Written by Himself* (1777).

Recommended modern editions of separate works by Hume include: *A Treatise of Human Nature,* ed. by L.A. Selby-Bigge, 2nd ed. rev. by P.H. Nidditch (1980); *Enquiries Concerning Human Understanding and Concerning the Principles of Morals,* ed. by L.A. Selby-Bigge, 3rd ed. rev. by P.H. Nidditch (1975); *Dialogues Concerning Natural Religion,* ed. by Norman Kemp Smith (1935, reissued 1981); *The Natural History of Religion,* ed. by A. Wayne Colver (1976); and *The History of Great Britain: The Reigns of James I and Charles I,* ed. by Duncan Forbes (1970). The best collected edition is *The Philosophical Works of David Hume,* ed. by T.H. Green and T.H. Grose, new ed., 4 vol. (1882–86, reprinted 1964).

BIBLIOGRAPHY

Biographies: The standard biography is ERNEST C. MOSSNER, *The Life of David Hume,* 2nd rev. ed. (1980); while JOHN HILL BURTON, *Life and Correspondence of David Hume,* 2 vol. (1846, reprinted 1983), is not entirely superseded by more recent scholarship. Shorter biographical studies include J.Y.T. GREIG, *David Hume* (1931, reprinted 1983); and F.H. HEINEMANN, *David Hume, the Man and His Science of Man* (1940). J.Y.T. GREIG (ed.), *The Letters of David Hume,* 2 vol. (1932, reprinted 1983), is supplemented by RAYMOND KLIBANSKY and ERNEST C. MOSSNER (eds.), *New Letters of David Hume* (1954, reprinted 1983).

Commentaries: An up-to-date introduction to Hume's philosophical ideas is provided by A.J. AYER, *Hume* (1980). Other useful introductory works are D.G.C. MACNABB, *David Hume: His Theory of Knowledge and Morality,* 2nd ed. (1966); BARRY STROUD, *Hume* (1977); and V.C. CHAPPELL (ed.), *The Philosophy*

of David Hume (1963), which contains excerpts from Hume's works. More specialized commentaries on Hume's epistemological theory include H.H. PRICE, *Hume's Theory of the External World* (1940, reprinted 1981); CONSTANCE MAUND, *Hume's Theory of Knowledge* (1937, reprinted 1973); and NORMAN KEMP SMITH, *The Philosophy of David Hume* (1941, reprinted 1983). Commentaries that concentrate on Hume's ethical thought are R. DAVID BROILES, *The Moral Philosophy of David Hume*, 2nd ed. (1969); J.L. MACKIE, *Hume's Moral Theory* (1980); ANTONY FLEW, *Hume's Philosophy of Belief* (1961); RACHEL M. KYDD, *Reason and Conduct in Hume's Treatise* (1946, reissued 1964); and JONATHAN HARRISON, *Hume's Moral Epistemology* (1976), and *Hume's Theory of Justice* (1981, reprinted 1983). ALASDAIR MACINTYRE (ed.), *Hume's Ethical Writings* (1965, reprinted 1979), contains excerpts that are well chosen from various works.

Hume's political theory is examined in DUNCAN FORBES, *Hume's Philosophical Politics* (1975, reprinted 1985); DAVID MILLER, *Philosophy and Ideology in Hume's Political Thought* (1981, reprinted 1984); G.K. VLACHOS, *Essai sur la politique de Hume* (1955); and SHIRLEY ROBIN LETWIN, *The Pursuit of Certainty: David Hume, Jeremy Bentham, John Stuart Mill, Beatrice Webb* (1965). Excerpts from Hume's writings on politics, together with a commentary, appear in FREDERICK WATKINS (ed.), *Theory of Politics* (1951).

RICHARD WOLLHEIM (ed.), *Hume on Religion* (1963, reissued 1964), contains Hume's most important writings on the subject. More extensive commentaries are provided by J.C.A. GASKIN, *Hume's Philosophy of Religion* (1978); and ANDRÉ-LOUIS LEROY, *La Critique et la religion chez David Hume* (1930).

EUGENE ROTWEIN (ed.), *Writings on Economics* (1955, reissued 1972), a collection of Hume's essays, is a valuable introduction to a somewhat neglected aspect of Hume's social theory. Commentaries that attempt a more synoptic and comprehensive appraisal of Hume's thought include JOHN PASSMORE, *Hume's Intentions*, 3rd ed. (1980); CHARLES W. HENDEL, *Studies in the Philosophy of David Hume* (1925, reissued 1983); ANDRÉ-LOUIS LEROY, *David Hume* (1953, reprinted 1979); JOHN LAIRD, *Hume's Philosophy of Human Nature* (1932, reprinted 1983); A.P. CAVENDISH (pseudonym A.H. BASSON), *David Hume* (1958, reprinted 1981); FARHANG ZABEEH, *Hume, Precursor of Modern Empiricism*, 2nd rev. ed. (1973); and JAMES NOXON, *Hume's Philosophical Development* (1973, reprinted 1975).

Bibliographies: T.E. JESSOP, *A Bibliography of David Hume and of Scottish Philosophy from Frances Hutchinson to Lord Balfour* (1938, reprinted 1983); and ROLAND HALL, *Fifty Years of Hume Scholarship* (1978), and *A Hume Bibliography, from 1930* (1971). WILLIAM B. TODD (ed.), *Hume and the Enlightenment* (1974), contains a bibliographical essay. *Hume Studies* (semiannual) updates bibliographical information.

(T.E.Je./M.C.)

Humour and Wit

In all its many-splendoured varieties, humour can be simply defined as a type of stimulation that tends to elicit the laughter reflex. Spontaneous laughter is a motor reflex produced by the coordinated contraction of 15 facial muscles in a stereotyped pattern and accompanied by altered breathing. Electrical stimulation of the main lifting muscle of the upper lip, the zygomatic major, with currents of varying intensity produces facial expressions ranging from the faint smile through the broad grin to the contortions typical of explosive laughter.

The laughter and smile of civilized man is, of course, often of a conventional kind, in which voluntary intent substitutes for, or interferes with, spontaneous reflex activity; this article is concerned, however, only with the latter. Once laughter is realized to be a humble reflex, several paradoxes must be faced. Motor reflexes, such as the contraction of the pupil of the eye in dazzling light, are simple responses to simple stimuli whose value to survival is obvious. But the involuntary contraction of 15 facial muscles, associated with certain irrepressible noises, strikes one as an activity without any utilitarian value, quite unrelated to the struggle for survival. Laughter is a reflex but unique in that it has no apparent biological purpose. One might call it a luxury reflex. Its only function seems to be to provide relief from tension.

The second related paradox is a striking discrepancy between the nature of the stimulus and that of the response in humorous transactions. When a blow beneath the kneecap causes an automatic upward kick, both "stimulus" and "response" function on the same primitive physiological level, without requiring the intervention of the higher mental functions. But that such a complex mental activity as reading a comic story should cause a specific reflex contraction of the facial muscles is a phenomenon that has puzzled philosophers since Plato. There is no clear-cut, predictable response that would tell a lecturer whether he has succeeded in convincing his listeners; but, when he is telling a joke, laughter serves as an experimental test. *Humour is the only form of communication in which a stimulus on a high level of complexity produces a stereotyped, predictable response on the physiological reflex level.* Thus the response can be used as an indicator for the presence of the elusive quality that is called humour—as the click of the Geiger counter is used to indicate the presence of radioactivity. Such a procedure is not possible in any other form of art; and, since the step from the sublime to the ridiculous is reversible, the study of humour provides clues for the study of creativity in general.

This article deals with the changing concepts and practice of humour from the time of Aristotle to the influence of the mass media in the contemporary world.

The article is divided into the following sections:

The logic of laughter 682
Laughter and emotion 683
Verbal humour 684
Situational humour 685
Styles and techniques in humour 686
Relations to art and science 686
The humanization of humour 687
Non-Western styles
Humour in the contemporary world
Bibliography 687

THE LOGIC OF LAUGHTER

The range of laughter-provoking experiences is enormous, from physical tickling to mental titillations of the most varied kinds. There is unity in this variety, however, a common denominator of a specific and specifiable pattern that reflects the "logic" or "grammar" of humour, as it were. A few examples will help to unravel that pattern.

1. A masochist is a person who likes a cold shower in the morning so he takes a hot one.

2. An English lady, on being asked by a friend what she thought of her departed husband's whereabouts: "Well, I suppose the poor soul is enjoying eternal bliss, but I wish you wouldn't talk about such unpleasant subjects."

3. A doctor comforts his patient: "You have a very serious disease. Of 10 persons who catch it, only one survives. It is lucky you came to me, for I have recently had nine patients with this disease and they all died of it."

4. Dialogue in a French film:
"Sir, I would like to ask for your daughter's hand."
"Why not? You have already had the rest."

5. A marquis of the court of Louis XV unexpectedly returned from a journey and, on entering his wife's boudoir, found her in the arms of a bishop. After a moment's hesitation, the marquis walked calmly to the window, leaned out, and began going through the motions of blessing the people in the street.
"What are you doing?" cried the anguished wife.
"Monseigneur is performing my functions, so I am performing his."

Is there a common pattern underlying these five stories? Starting with the last, a little reflection reveals that the marquis's behaviour is both unexpected and perfectly logical—but of a logic not usually applied to this type of situation. It is the logic of the division of labour, governed by rules as old as human civilization. But his reactions would have been expected to be governed by a different set of rules—the code of sexual morality. It is the sudden clash between these two mutually exclusive codes of rules—or associative contexts—that produces the comic effect. It compels the listener to perceive the situation in two self-consistent but incompatible frames of reference at the same time; his mind has to operate simultaneously on two different wavelengths. While this unusual condition lasts, the event is not only, as is normally the case, associated with a single frame of reference but "disociated" with two. The word disociation was coined by the present writer to make a distinction between the routines of disciplined thinking within a single universe of discourse—on a single plane, as it were—and the creative types of mental activity that always operate on more than one plane. In humour, both the *creation* of a subtle joke and the *re-creative* act of perceiving the joke involve the delightful mental jolt of a sudden leap from one plane or associative context to another. Disociation

Turning to the other examples, in the French film dialogue, the daughter's "hand" is perceived first in a metaphorical frame of reference, then suddenly in a literal, bodily context. The doctor thinks in terms of abstract, statistical probabilities, the rules of which are inapplicable to individual cases; and there is an added twist because, in contrast to what common sense suggests, the patient's odds of survival are unaffected by whatever happened before; they are still one against 10. This is one of the profound paradoxes of the theory of probability, and the joke in fact implies a riddle; it pinpoints an absurdity that tends to be taken for granted. As for the lady who looks upon death as "eternal bliss" and at the same time "an unpleasant subject," she epitomizes the common human predicament of living in the divided house of faith and reason. Here again the simple joke carries unconscious overtones and undertones, audible to the inner ear alone.

The masochist who punishes himself by depriving himself of his daily punishment is governed by rules that are a *reversal* of those of normal logic. (A pattern can be constructed in which *both* frames of reference are reversed: "A sadist is a person who is kind to a masochist.") But there

is again an added twist. The joker does not really believe that the masochist takes his hot shower as a punishment; he only pretends to believe it. *Irony* is the satirist's most effective weapon; it pretends to adopt the opponent's ways of reasoning in order to expose their implicit absurdity or viciousness.

The common pattern underlying these stories is *the perceiving of a situation in two self-consistent but mutually incompatible frames of reference or associative contexts.* This formula can be shown to have a general validity for all forms of humour and wit, some of which will be discussed below. But it covers only one aspect of humour—its *intellectual structure.* Another fundamental aspect must be examined—the *emotional dynamics* that breathe life into that structure and make a person laugh, giggle, or smile.

LAUGHTER AND EMOTION

When a comedian tells a story, he deliberately sets out to create a certain tension in his listeners, which mounts as the narrative progresses. But it never reaches its expected climax. The punch line, or point, acts as a verbal guillotine that cuts across the logical development of the story; it debunks the audience's dramatic expectations. The tension that was felt becomes suddenly redundant and is exploded in laughter. To put it differently, laughter disposes of emotive excitations that have become pointless and must somehow be worked off along physiological channels of least resistance; and the function of the "luxury reflex" is to provide these channels.

A glance at the caricatures of the 18th-century English artists William Hogarth or Thomas Rowlandson, showing the brutal merriment of people in a tavern, makes one realize at once that they are working off their surplus of adrenalin by contracting their face muscles into grimaces, slapping their thighs, and breathing in puffs through the half-closed glottis. Their flushed faces reveal that the emotions disposed of through these safety valves are brutality, envy, sexual gloating. In cartoons by the 20th-century American James Thurber, however, coarse laughter yields to an amused and rarefied smile: the flow of adrenalin has been distilled and crystallized into a grain of Attic salt—a sophisticated joke. The word witticism is derived from "wit" in its original sense of intelligence and acumen (as is *Witz* in German). The domains of humour and of ingenuity are continuous, without a sharp boundary: the jester is brother to the sage. Across the spectrum of humour, from its coarse to its subtle forms, from practical joke to brainteaser, from jibe to irony, from anecdote to epigram, the emotional climate shows a gradual transformation. The emotion discharged in coarse laughter is aggression robbed of its purpose. The jokes small children enjoy are mostly scatological; adolescents of all ages gloat on vicarious sex. The sick joke trades on repressed sadism, satire on righteous indignation. There is a bewildering variety of moods involved in different forms of humour, including mixed or contradictory feelings; but whatever the mixture, it must contain a basic ingredient that is indispensable: an impulse, however faint, of aggression or apprehension. It may appear in the guise of malice, contempt, the veiled cruelty of condescension, or merely an absence of sympathy with the victim of the joke—a momentary anesthesia of the heart, as the French philosopher Henri Bergson put it.

The aggressive element

In the subtler types of humour, the aggressive tendency may be so faint that only careful analysis will detect it, like the presence of salt in a well-prepared dish—which, however, would be tasteless without it. Replace aggression by sympathy and the same situation—a drunk falling on his face, for example—will be no longer comic but pathetic and will evoke not laughter but pity. It is the aggressive element, the detached malice of the comic impersonator, that turns pathos into bathos, tragedy into travesty. Malice may be combined with affection in friendly teasing; and the aggressive component in civilized humour may be sublimated or no longer conscious. But in jokes that appeal to children and primitive people, cruelty and boastful self-assertiveness are much in evidence. In 1961 a survey carried out among American children aged eight to 15 made the researchers conclude that the mortification, discomfort, or hoaxing of others readily caused laughter, but witty or funny remarks often passed unnoticed.

Theories of laughter

Similar considerations apply to the historically earlier forms and theories of the comic. In Aristotle's view, laughter was intimately related to ugliness and debasement. Cicero held that the province of the ridiculous lay in a certain baseness and deformity. Descartes believed that laughter was a manifestation of joy mixed with surprise or hatred or both. In Francis Bacon's list of what causes laughter, the first place is again given to deformity. One of the most frequently quoted utterances on the subject is this definition in Thomas Hobbes's *Leviathan* (1651):

> The passion of laughter is nothing else but sudden glory arising from a sudden conception of some eminency in ourselves by comparison with the infirmity of others, or with our own formerly.

In the 19th century, Alexander Bain, an early experimental psychologist, thought along the same lines:

> Not in physical effects alone, but in everything where a man can achieve a stroke of superiority, in surpassing or discomforting a rival, is the disposition of laughter apparent.

In Bergson's view, laughter is the corrective punishment inflicted by society upon the unsocial individual: "In laughter we always find an unavowed intention to humiliate and consequently to correct our neighbour." Sir Max Beerbohm, the 20th-century English wit, found "two elements in the public's humour: delight in suffering, contempt for the unfamiliar." The American psychologist William McDougall believed that "laughter has been evolved in the human race as an antidote to sympathy, a protective reaction shielding us from the depressive influence of the shortcomings of our fellow men."

However much the opinions of the theorists differ, on this one point nearly all of them agree: that the emotions discharged in laughter always contain an element of aggressiveness. It must be borne in mind, however, that aggression and apprehension are twin phenomena, so much so that psychologists are used to talking of "aggressive–defensive impulses." Accordingly, one of the typical situations in which laughter occurs is the moment of sudden cessation of fear caused by some imaginary danger. Rarely is the nature of laughter as an overflow of redundant tensions more strikingly manifested than in the sudden change of expression on a small child's face from anxious apprehension to the happy laughter of relief. This seems to be unrelated to humour; yet a closer look reveals in it the same logical structure as in the joke: the wildly barking little dog was first perceived by the child in a context of danger, then discovered to be a harmless pup; the tension has suddenly become redundant and is spilled.

Laughter as the release of tension

Immanuel Kant realized that what causes laughter is "the sudden transformation of a tense expectation into nothing." Herbert Spencer, the 19th-century English philosopher, took up the idea and attempted to formulate it in physiological terms: "Emotions and sensations tend to generate bodily movements. . . . When consciousness is unawares transferred from great things to small," the "liberated nerve force" will expend itself along channels of least resistance—the bodily movements of laughter. Freud incorporated Spencer's theory of humour into his own, with special emphasis on the release of repressed emotions in laughing; he also attempted to explain why the excess energy should be worked off in that particular way:

> According to the best of my knowledge, the grimaces and contortions of the corners of the mouth that characterise laughter appear first in the satisfied and over-satiated nursling when he drowsily quits the breast. . . . They are physical expressions of the determination to take no more nourishment, an "enough" so to speak, or rather a "more than enough" . . . This primal sense of pleasurable saturation may have provided the link between the smile—that basic phenomenon underlying laughter—and its subsequent connection with other pleasurable processes of de-tension.

In other words, the muscle contractions of the smile, as the earliest expressions of relief from tension, would thereafter serve as channels of least resistance. Similarly, the explosive exhalations of laughter seem designed to "puff away" surplus tension in a kind of respiratory gymnastics, and agitated gestures obviously serve the same function.

It may be objected that such massive reactions often seem quite out of proportion to the slight stimulations that provoke them. But it must be borne in mind that laughter is a phenomenon of the trigger-releaser type, where a sudden turn of the tap may release vast amounts of stored emotions, derived from various, often unconscious, sources: repressed sadism, sexual tumescence, unavowed fear, even boredom. The explosive laughter of a class of schoolboys at some trivial incident is a measure of their pent-up resentment during a boring lecture. Another factor that may amplify the reaction out of all proportion to the comic stimulus is the social infectiousness that laughter shares with other emotive manifestations of group behaviour.

Patterns of association

Laughter or smiling may also be caused by stimulations that are not in themselves comic but signs or symbols deputizing for well-established comic patterns—such as Charlie Chaplin's oversized shoes or Groucho Marx's cigar—or catchphrases, or allusions to family jokes. To discover why people laugh requires, on some occasions, tracing back a long, involved thread of associations to its source. This task is further complicated by the fact that the effect of such comic symbols—in a cartoon or on the stage—appears to be instantaneous, without allowing time for the accumulation and subsequent discharge of "expectations" and "emotive tensions." But here memory comes into play, having already accumulated the required emotions in past experiences, acting as a storage battery whose charge can be sparked off at any time: the smile that greets Falstaff's appearance on the scene is derived from a mixture of memories and expectations. Besides, even if a reaction to a cartoon appears to be instantaneous, there is always a process in time until the reader "sees the joke"; the cartoon has to tell a story even if it is telescoped into a few seconds. All of this shows that to analyze humour is a task as delicate as analyzing the composition of a perfume with its multiple ingredients, some of which are never consciously perceived while others, when sniffed in isolation, would make one wince.

In this article there has been a discussion first of the logical structure of humour and then of its emotional dynamics. Putting the two together, the result may be summarized as follows: the "bisociation" of a situation or idea with two mutually incompatible contexts in a person's mind and the resulting abrupt transfer of his train of thought from one context to another put a sudden end to his "tense expectations"; the accumulated emotion, deprived of its object, is left hanging in the air and is discharged in laughter. Upon hearing that the marquis in the story told earlier walks to the window and starts blessing the people in the street, the intellect turns a somersault and enters with gusto into the new game. The malicious and erotic feelings aroused by the start of the story, however, cannot be fitted into the new context; deserted by the nimble intellect, these feelings gush out in laughter like air from a punctured tire.

Separation of thought and emotion

To put it differently: people laugh because their emotions have a greater inertia and persistence than their thoughts. Affects are incapable of keeping step with reasoning; unlike reasoning, they cannot "change direction" at a moment's notice. To the physiologist, this is self-evident since emotions operate through the genetically old, massive sympathetic nervous system and its allied hormones, acting on the whole body, while the processes of conceptual thinking are confined to the neocortex at the roof of the brain. Common experience provides daily confirmation of this dichotomy. People are literally "poisoned" by their adrenal humours; it takes time to talk a person out of a mood; fear and anger show physical aftereffects long after their causes have been removed. If man were able to change his moods as quickly as his thoughts, he would be an acrobat of emotion; but since he is not, his thoughts and emotions frequently become dissociated. It is emotion deserted by thought that is discharged in laughter. For emotion, owing to its greater mass momentum, is, as has been shown, unable to follow the sudden switch of ideas to a different type of logic; it tends to persist in a straight line. Aldous Huxley once wrote:

> We carry around with us a glandular system which was admirably well adapted to life in the Paleolithic times but is not very well adapted to life now. Thus we tend to produce more adrenalin than is good for us, and we either suppress ourselves and turn destructive energies inwards or else we do not suppress ourselves and we start hitting people. (From *Man and Civilization: Control of the Mind*, ed. Seymour M. Farber and Roger H.L. Wilson. Copyright 1961. Used with permission of McGraw-Hill Book Company.)

A third alternative is to laugh at people. There are other outlets for tame aggression, such as competitive sports or literary criticism; but they are acquired skills, whereas laughter is a gift of nature, included in man's native equipment. The glands that control his emotions reflect conditions at a stage of evolution when the struggle for existence was more deadly than at present—and when the reaction to any strange sight or sound consisted in jumping, bristling, fighting, or running. As security and comfort increased in the species, new outlets were needed for emotions that could no longer be worked off through their original channels, and laughter is obviously one of them. But it could only emerge when reasoning had gained a degree of independence from the urges of emotion. Below the human level, thinking and feeling appear to form an indivisible unity. Not before thinking became gradually detached from feeling could man perceive his own emotion as redundant and make the smiling admission, "I have been fooled."

VERBAL HUMOUR

The foregoing discussion was intended to provide the tools for dissecting and analyzing any specimen of humour. The procedure is to determine the nature of the two (or more) frames of reference whose collision gives rise to the comic effect—to discover the type of logic or "rules of the game" that govern each. In the more sophisticated type of joke, the logic is implied and hidden, and the moment it is stated in explicit form, the joke is dead. Unavoidably, the section that follows will be strewn with cadavers.

Puns

Max Eastman, in *The Enjoyment of Laughter* (1936), remarked of a laboured pun by Ogden Nash: "It is not a pun but a punitive expedition." That applies to most puns, including Milton's famous lines about the Prophet Elijah's ravens, which were "though ravenous taught to abstain from what they brought," or the character mentioned by Freud, who calls the Christmas season the "alcoholidays." Most puns strike one as atrocious, perhaps because they represent the most primitive form of humour; two disparate strings of thought tied together by an acoustic knot. But the very primitiveness of such association based on pure sound ("hol") may account for the pun's immense popularity with children and its prevalence in certain types of mental disorder ("punning mania").

From the play on sounds—puns and Spoonerisms—an ascending series leads to the play on words and so to the play on ideas. When Groucho Marx says of a safari in Africa, "We shot two bucks, but that was all the money we had," the joke hinges on the two meanings of the word buck. It would be less funny without the reference to Groucho, which evokes a visual image instantly arousing high expectations. The story about the marquis above may be considered of a superior type of humour because it plays not on mere words but on ideas.

It would be quite easy—and equally boring—to draw up a list in which jokes and witticisms are classified according to the nature of the frames of reference whose collision creates the comic effect. A few have already been mentioned: metaphorical versus literal meaning (the daughter's "hand"); professional versus common sense logic (the doctor); incompatible codes of behaviour (the marquis); confrontations of the trivial and the exalted ("eternal bliss"); trains of reasoning travelling, happily joined together, in opposite directions (the sadist who is kind to the masochist). The list could be extended indefinitely; in fact *any* two frames of reference can be made to yield a comic effect of sorts by hooking them together and infusing a drop of malice into the concoction. The frames may even be defined by such abstract concepts as "time" and "weather": the absent-minded professor who tries to read the temperature from his watch or to tell the time from the thermometer is comic in the same way as a game of

table tennis played with a soccer ball or a game of rugby played with a table tennis ball. The variations are infinite, the formula remains the same.

Jokes and anecdotes have a single point of culmination. The literary forms of *sustained humour*, such as the picaresque novel, do not rely on a single effect but on a series of minor climaxes. The narrative moves along the line of intersection of contrasted planes, such as the fantasy world of Don Quixote and the cunning horse sense of Sancho Panza, or is made to oscillate between them. As a result, tension is continuously generated and discharged in mild amusement.

Comic verse

Comic verse thrives on the melodious union of incongruities, such as the "cabbages and kings" in Lewis Carroll's "The Walrus and the Carpenter," and particularly on the contrast between lofty form and flat-footed content. Certain metric forms associated with heroic poetry, such as the hexameter or Alexandrine, arouse expectations of pathos, of the exalted; to pour into these epic molds some homely, trivial content—"beautiful soup, so rich and green/ waiting in a hot tureen"—is an almost infallible comic device. The rolling rhythms of the first lines of a limerick that carry, instead of a mythical hero such as Hector or Achilles, a young lady from Ohio for a ride make her ridiculous even before the expected calamities befall her. Instead of a heroic mold, a soft lyrical one may also pay off:

> . . . And what could be moister
> Than tears of an oyster?

Another type of incongruity between form and content yields the bogus proverb: "The rule is: jam tomorrow and jam yesterday—but never jam today." Two contradictory statements have been telescoped into a line whose homely, admonitory sound conveys the impression of a popular adage. In a similar way, nonsense verse achieves its effect by pretending to make sense, by forcing the reader to project meaning into the phonetic pattern of the jabberwocky, as one interprets the ink blots in a Rorschach test.

Satire and allegory

The *satire* is a verbal caricature that shows a deliberately distorted image of a person, institution, or society. The traditional method of the caricaturist is to exaggerate those features he considers to be characteristic of his victim's personality and to simplify by leaving out everything that is not relevant for his purpose. The satirist uses the same technique, and the features of society he selects for magnification are, of course, those of which he disapproves. The result is a juxtaposition, in the reader's mind, of his habitual image of the world in which he moves and its absurd reflection in the satirist's distorting mirror. He is made to recognize familiar features in the absurd and absurdity in the familiar. Without this double vision the satire would be humourless. If the human Yahoos were really such evil-smelling monsters as Gulliver's Houyhnhnm hosts claim, then Jonathan Swift's *Gulliver's Travels* (1726) would not be a satire but the statement of a deplorable truth. Straight invective is not satire; satire must deliberately overshoot its mark.

A similar effect is achieved if, instead of exaggerating the objectionable features, the satirist projects them by means of the *allegory* onto a different background, such as an animal society. A succession of writers, from the ancient Greek dramatist Aristophanes through Swift to such 20th-century satirists as Anatole France and George Orwell, have used this technique to focus attention on deformities of society that, blunted by habit, are taken for granted.

SITUATIONAL HUMOUR

The coarsest type of humour is the practical joke: pulling away the chair from under the dignitary's lowered bottom. The victim is perceived first as a person of consequence, then suddenly as an inert body subject to the laws of physics: authority is debunked by gravity, mind by matter; man is degraded to a mechanism. Goose-stepping soldiers acting like automatons, the pedant behaving like a mechanical robot, the Sergeant Major attacked by diarrhea, or Hamlet getting the hiccups—all show man's lofty aspirations deflated by his all-too-solid flesh. A similar effect is produced by artifacts that masquerade as humans: Punch and Judy, jack-in-the-box, gadgets playing tricks on their masters as if with calculated malice.

In Henri Bergson's theory of laughter, this dualism of subtle mind and inert matter—he calls it "the mechanical encrusted on the living"—is made to serve as an explanation of *all* varieties of the comic. In the light of what has been said, however, it would seem to apply only to *one* type of comic situation among many others.

From the "bisociation" of man and machine, there is only a step to the man–animal hybrid. Walt Disney's creations behave as if they were human without losing their animal appearance. The caricaturist follows the reverse procedure by discovering horsey, mousy, or piggish features in the human face.

This leads to the comic devices of imitation, impersonation, and disguise. The impersonator is perceived as himself and somebody else at the same time. If the result is slightly degrading—but only in that case—the spectator will laugh. The comedian impersonating a public personality, two pairs of trousers serving as the legs of the pantomime horse, men disguised as women and women as men—in each case the paired patterns reduce each other to absurdity.

The most aggressive form of impersonation is the *parody*, designed to deflate hollow pretense, to destroy illusion, and to undermine pathos by harping on the weaknesses of the victim. Wigs falling off, speakers forgetting their lines, gestures remaining suspended in the air: the parodist's favourite points of attack are again situated on the line of intersection between the sublime and the trivial.

Playful behaviour in young animals and children is amusing because it is an unintentional parody of adult behaviour, which it imitates or anticipates. Young puppies are droll because their helplessness, affection, and puzzled expression make them appear more "human" than full-grown dogs; because their growls strike one as impersonations of adult behaviour—like a child in a bowler hat; because the puppy's waddling, uncertain gait makes it a choice victim of nature's practical jokes; because its bodily disproportions—the huge padded paws, Falstaffian belly, and wrinkled brow—give it the appearance of a caricature; and lastly because the observer feels so superior to a puppy. A fleeting smile can contain many logical ingredients and emotional spices.

Both Cicero and Francis Bacon regarded *deformity* as the most frequent cause of laughter. Renaissance princes collected dwarfs and hunchbacks for their merriment. It obviously requires a certain amount of imagination and empathy to recognize in a midget a fellow human, who, though different in appearance, thinks and feels much as oneself does. In children, this projective faculty is still rudimentary: they tend to mock people with a stammer or a limp and laugh at the foreigner with an odd pronunciation. Similar attitudes are shown by tribal or parochial societies to any form of appearance or behaviour that deviates from their strict norms: the stranger is not really human; he only pretends to be "like us." The Greeks used the same word, barbarous, for the foreigner and the stutterer: the uncouth barking sounds the stranger uttered were considered a parody of human speech. Vestiges of this primitive attitude are still found in the curious fact that civilized people accept a foreign accent with tolerance, whereas imitation of a foreign accent strikes them as comic. The imitator's mispronunciations are recognized as mere pretense; this knowledge makes sympathy unnecessary and enables the audience to be childishly cruel with a clean conscience.

Other sources of innocent laughter are situations in which *the part and the whole* change roles, and attention becomes focussed on a detail torn out of the functional context on which its meaning depended. When the phonograph needle gets stuck, the soprano's voice keeps repeating the same word on the same quaver, which suddenly assumes a grotesquely independent life. The same happens when faulty orthography displaces attention from meaning to spelling, or whenever consciousness is directed at functions that otherwise are performed automatically. The latter situation is well illustrated by the story of the centipede who, when asked in which order he moved his

hundred legs, became paralyzed and could walk no more. The self-conscious, awkward youth, who does not know what to do with his hands, is a victim of the paradox of the centipede.

Comedies have been classified according to their reliance on situations, manners, or characters. The logic of the last two needs no further discussion; in the first, comic effects are contrived by making a situation participate simultaneously in two independent chains of events with different associative contexts, which intersect through coincidence, mistaken identity, or confusions of time and occasion.

Tickling

Why tickling should produce laughter remained an enigma in all earlier theories of the comic. As Darwin was the first to point out, the innate response to tickling is squirming and straining to withdraw the tickled part—a defense reaction designed to escape attacks on vulnerable areas such as the soles of the feet, armpits, belly, and flank. If a fly settles on the belly of a horse, it causes a ripple of muscle contractions across the skin—the equivalent of squirming in the tickled child. But the horse does not laugh when tickled, and the child not always. The child will laugh only—and this is the crux of the matter—when it perceives tickling as a *mock attack,* a caress in mildly aggressive disguise. For the same reason, people laugh only when tickled by others, not when they tickle themselves.

Experiments at Yale University on babies under one year revealed the not very surprising fact that they laughed 15 times more often when tickled by their mothers than by strangers; and when tickled by strangers, they mostly cried. For the mock attack must be recognized as being only pretense, and with strangers one cannot be sure. Even with its own mother, there is an ever-so-slight feeling of uncertainty and apprehension, the expression of which will alternate with laughter in the baby's behaviour. It is precisely this element of tension between the tickles that is relieved in the laughter accompanying the squirm. The rule of the game is "let me be just a little frightened so that I can enjoy the relief."

Thus the tickler is impersonating an aggressor but is simultaneously known not to be one. This is probably the first situation in life that makes the infant live on two planes at once, a delectable foretaste of being tickled by the horror comic.

Humour in the visual arts and music

Humour in the visual arts reflects the same logical structures as discussed before. Its most primitive form is the distorting mirror at the fun fair, which reflects the human frame elongated into a column or compressed into the shape of a toad. It plays a practical joke on the victim, who sees the image in the mirror both as his familiar self and as a lump of plasticine that can be stretched and squeezed into any absurd form. The mirror distorts mechanically while the caricaturist does so selectively, employing the same method as the satirist—exaggerating characteristic features and simplifying the rest. Like the satirist, the caricaturist reveals the absurd in the familiar; and, like the satirist, he must overshoot his mark. His malice is rendered harmless by the knowledge that the monstrous potbellies and bowlegs he draws are not real; real deformities are not comic but arouse pity.

The artist, painting a stylized portrait, also uses the technique of selection, exaggeration, and simplification; but his attitude toward the model is usually dominated by positive empathy instead of negative malice, and the features he selects for emphasis differ accordingly. In some character studies by Leonardo da Vinci, Hogarth, or Honoré Daumier, the passions reflected are so violent, the grimaces so ferocious, that it is impossible to tell whether the works were meant as portraits or caricatures. If one feels that such distortions of the human face are not really possible, that Daumier merely *pretended* that they exist, then one is absolved from horror and pity and can laugh at his grotesques. But if one feels that this is indeed what Daumier saw in those dehumanized faces, then they are not comic but tragic.

Humour in *music* is a subject to be approached with diffidence because the language of music ultimately eludes translation into verbal concepts. All one can do is to point out some analogies: a "rude" noise, such as the blast of a trumpet inserted into a passage where it does not belong, has the effect of a practical joke; a singer or an instrument out of tune produces a similar reaction; the imitation of animal sounds, vocally or instrumentally, exploits the technique of impersonation; a nocturne by Chopin transposed into hot jazz or a simple street song performed with Wagnerian pathos is a marriage of incompatibles. These are primitive devices corresponding to the lowest levels of humour; more sophisticated are the techniques employed by Maurice Ravel in *La Valse,* a parody of the sentimental Viennese waltz, or by Zoltán Kodály in the mock-heroics of his Hungarian folk opera, *Háry János.* But in comic operas it is almost impossible to sort out how much of the comic effect is derived from the book and how much from the music; and the highest forms of musical humour, the unexpected delights of a lighthearted scherzo by Mozart, defy verbal analysis, unless it is so specialized and technical as to defeat its purpose. Although a "witty" musical passage that springs a surprise on the audience and cheats it of its expectations certainly has the emotion-relieving effect that tends to produce laughter, a concert audience may occasionally smile but will hardly ever laugh: the emotions evoked by musical humour are of a subtler kind than those of the verbal and visual variety.

STYLES AND TECHNIQUES IN HUMOUR

The criteria that determine whether a humorous offering will be judged good, bad, or indifferent are partly a matter of period taste and personal preference and partly dependent on the style and technique of the humorist. It would seem that these criteria can be summed up under three main headings: originality, emphasis, and economy.

Originality, emphasis, and economy

The merits of originality are self-evident; it provides the essential element of surprise, which cuts across our expectations. But true originality is not very often met either in humour or in other forms of art. One common substitute for it is to increase the tension of the audience by various techniques of suggestive emphasis. The clown's domain is the rich, coarse type of humour: he piles it on; he appeals to sadistic, sexual, scatological impulses. One of his favourite tricks is repetition of the same situation, the same key phrase. This diminishes the effect of surprise, but it has a tension-accumulating effect: emotion is easily drawn into the familiar channel—more and more liquid is being pumped into the punctured pipeline.

Emphasis on local colour and ethnic peculiarities, such as Scottish or Cockney stories, for example, is a further means to channel emotion into familiar tracks. The Scotsman or Cockney must, of course, be a caricature if the comic purpose is to be achieved. In other words, exaggeration and simplification once more appear as indispensable tools to provide emphasis.

In the higher forms of humour, however, emphasis tends to yield to the opposite kind of virtue—economy. Economy, in humour and art, does not mean mechanical brevity but implicit hints instead of explicit statements—the oblique allusion in lieu of the frontal attack. Old-fashioned cartoons, such as those featuring the British lion and the Russian bear, hammered their message in; the modern cartoon usually poses a riddle that the reader must solve by an imaginative effort in order to see the joke.

In humour, as in other forms of art, emphasis and economy are complementary techniques. The first forces the offering down the consumer's throat; the second tantalizes to whet his appetite.

RELATIONS TO ART AND SCIENCE

Earlier theories of humour, including even those of Bergson and Freud, treated it as an isolated phenomenon, without attempting to throw light on the intimate connections between the comic and the tragic, between laughter and crying, between artistic inspiration, comic inventiveness, and scientific discovery. Yet these three domains of creative activity form a continuum with no sharp boundaries between wit and ingenuity, nor between discovery and art.

It has been said that scientific discovery consists in seeing an analogy where nobody has seen one before. When, in the Song of Solomon, Solomon compared the Shulamite's neck to a tower of ivory, he saw an analogy that nobody had seen before; when William Harvey compared

the heart of a fish to a mechanical pump, he did the same; and when the caricaturist draws a nose like a cucumber, he again does just that. In fact, all the logical patterns discussed above, which constitute a "grammar" of humour, can also enter the service of art or discovery, as the case may be. The pun has structural equivalents in the rhyme and in word games, which range from crossword puzzles to the deciphering of the Rosetta Stone, the key to Egyptian hieroglyphic. The confrontation between diverse codes of behaviour may yield comedy, tragedy, or new psychological insights. The dualism of mind and inert matter is exploited by the practical joker but also provides one of the eternal themes of literature: man as a marionette on strings, manipulated by gods or chromosomes. The man–beast dichotomy is reflected by Walt Disney's cartoon character Donald Duck but also in Franz Kafka's macabre tale *The Metamorphosis* (1915) and in the psychologist's experiments with rats. The caricature corresponds not only to the artist's character portrait but also to the scientist's diagrams and charts, which emphasize the relevant features and leave out the rest.

Synthesis, juxtaposition, and collision

Contemporary psychology regards the conscious and unconscious processes underlying creativity in all domains as an essentially combinative activity—the bringing together of previously separate areas of knowledge and experience. The scientist's purpose is to achieve *synthesis;* the artist aims at a *juxtaposition* of the familiar and the eternal; the humorist's game is to contrive a *collision.* And as their motivations differ, so do the emotional responses evoked by each type of creativity: discovery satisfies the exploratory drive; art induces emotional catharsis; humour arouses malice and provides a harmless outlet for it. Laughter has been described as the "Haha reaction"; the discoverer's Eureka cry as the "Aha! reaction"; and the delight of the aesthetic experience as the "Ah . . . reaction." But the transitions from one to the other are continuous: witticism blends into epigram, caricature into portrait; and whether one considers architecture, medicine, chess, or cookery, there is no clear frontier where the realm of science ends and that of art begins: the creative person is a citizen of both. Comedy and tragedy, laughter and weeping, mark the extremes of a continuous spectrum, and a comparison of the physiology of laughter and weeping yields further clues to this challenging problem. Such considerations, however, lie beyond the terms of reference of the present article.

THE HUMANIZATION OF HUMOUR

The San (Bushmen) of the Kalahari desert of South West Africa/Namibia are among the oldest and most primitive inhabitants of the Earth. An anthropologist who made an exhaustive study of them provided a rare glimpse of prehistoric humour:

> On the way home we saw and shot a springbok, as there was no meat left in camp. The bullet hit the springbok in the stomach and partly eviscerated him, causing him to jump and kick before he finally died. The Bushmen thought that this was terribly funny and they laughed, slapping their thighs and kicking their heels to imitate the springbok, showing no pity at all, but then they regard animals with great detachment.

But the San remained "in good spirits, pleased with the amusement the springbok had given them." (From Elizabeth Marshall Thomas, *The Harmless People,* Alfred A. Knopf, New York, 1959.)

Obviously the San, like most primitive people, do not regard animals as sentient beings; the springbok's kicking in his agony appears to them funny because in their view the animal *pretends* to suffer pain like a human being, though it is incapable of such feelings. The ancient Greeks' attitude toward the stammering barbarian was similarly inspired by the conviction that he is not really human but only pretends to be. The ancient Hebrews' sense of humour seems to have been no less harsh: it has been pointed out that in the Old Testament there are 29 references to laughter, out of which 13 instances are linked with scorn, derision, mocking, and contempt and only two are born of joy.

As laughter emerged from antiquity, it was so aggressive that it has been likened to a dagger. It was in ancient Greece that the dagger was transformed into a quill, dripping with poison at first, then diluted and infused with delightfully lyrical and fanciful ingredients. The 5th century BC saw the first rise of humour into art, starting with parodies of Olympian heroics and soon reaching a peak, in some respects unsurpassed to this day, in the comedies of Aristophanes. From here onward, the evolution of humour in the Western world merges with the history of literature and art.

If the overall trend was toward the humanization of humour from primitive to sophisticated forms, there also have been ups and downs reflecting changes in political and cultural climate. George Orwell's satire of the 20th century, for example, is much more savage than that of Jonathan Swift in 18th-century England or of Voltaire in 18th-century France. If the Dark Ages produced works of humorous art, little of it has survived. And under the tyrannies of Hitler in Germany and of Stalin in the Soviet Union, humour was driven underground. Dictators fear laughter more than bombs.

Non-Western styles. About non-Western varieties of humour, the Westerner is tempted to repeat the middle-aged British matron's remark on watching Cleopatra rave and die on the stage: "How different, how very different from the home life of our dear Queen." Humour thrives only in its native climate, embedded in its native logic; when one does not know what to expect, one cannot be cheated of one's expectations. Hindu humour, for instance, as exemplified by the savage pranks played on humans by the monkey-god Hanuman, strikes the Westerner as particularly cruel, perhaps because the Hindu's approach to mythology is fundamentally alien to the Western mind.

Japanese humour

The humour of the Japanese, on the other hand, is, from the Western point of view, astonishingly mild and poetical, like weak, mint-flavoured tea:

> The boss of the monkeys ordered his thousands of henchmen to get the moon reflected in the water. They all tried various means but failed and were much troubled. One of the monkeys at last got the moon in the water and respectfully offered it to the boss, saying "This is what you asked for." The boss was delighted and praised him, saying, "What an exploit! You have distinguished yourself!" The monkey then asked, "By the way, master, what are you going to do with this?" "Well, yes . . . I didn't think of that." (From *Karukuchi Ukibyotan,* 1751; in R.H. Blyth, *Japanese Humour,* 1957.)

The following dates from about a century later:

> There was once a man who was always bewailing his lack of money to buy saké (rice wine) with. His wife, feeling sorry for him, dutifully cut off some of her hair and sold it to the hairdresser's for twenty-four mon, and bought her husband some saké. "Where on earth did you get this from?" "I sold my hair and bought it." "You did such a thing for me?" The wretched man shed tears, and fondling his wife's remaining hair said, "Yes, and there's another good half-bottle of saké here!" (From *Chanoko-mochi,* 1856; in Blyth.)

The combination of maudlin tears and brazen selfishness, and the crazy logic of equating the wife's coiffure with a liquid measure of saké, show the familiar Western pattern of the clash of incompatibles, even though transplanted into another culture.

Humour in the contemporary world. Humour today seems to be dominated by two main factors: the influence of the mass media and the crisis of values affecting a culture in rapid and violent transition. The former tends toward the commercialized manufacture of laughter by popular comedians and gags produced by conveyor-belt methods; the latter toward a sophisticated form of black humour larded with sick jokes, sadism, and sex.

Fashions, however, always run their course; perhaps the next one will delight in variations on the theme of the monkey boss who, having gained possession of the moon, does not know what to do with it. The only certainty regarding the humour of the future is contained in Dr. Samuel Johnson's dictum: "Sir, men have been wise in many different modes, but they have always laughed in the same way."

BIBLIOGRAPHY. GUILLAUME DUCHENNE (DE BOULOGNE), *Le Mécanisme de la physionomie humaine* (1862); and JULES-MARIE RAULIN, *Le Rire et les exhilarantes: étude anatomique, psycho-physiologique et pathologique* (1900), contain valuable

source material on the study of the physiology of laughter, including experiments by the neurologist Jean-Martin Charcot and the Nobel laureate Charles Richet. HERBERT SPENCER, in his "The Physiology of Laughter," *Essays on Education and Kindred Subjects* (1910, reprinted 1977), outlined the tension-relieving function of humour, on which FREUD elaborated in his *Wit and Its Relations to the Unconscious* (1916; originally published in German, 1905), with special emphasis on infantile and repressed elements. HENRI BERGSON, *Laughter* (1911, reprinted 1937; originally published in French, 1900), is a classic work attempting to derive all types of humour from the contrast between mind and matter. MAX F. EASTMAN, *Enjoyment of Laughter* (1936), is unique in that he denies the malicious element in laughter. REGINALD H. BLYTH, *Japanese Humor*, 2nd ed. (1961), throws a delightful sidelight on a different culture. DAVID H. MONRO, *Argument of Laughter* (1951, reprinted 1963), contains a valuable summary of earlier theories. ARTHUR KOESTLER, *Insight and Outlook* (1949, reprinted 1965), and *The Act of Creation* (1964, reprinted 1976), attempt to present a synthetic theory of humour and its interrelations with art and discovery.

(A.Ko.)

Hungarian Literature

No written evidence remains of the earliest Hungarian literature, but through Hungarian folktales and folk songs elements have survived that can be traced back to pagan times. Also extant, although only in Latin and dating from between the 11th and 14th centuries, are shortened versions of some Hungarian legends relating the origins of the Hungarian people and episodes from the conquest of Hungary and from the Hungarian campaigns of the 10th century.

This article is divided into the following sections:

Earliest writings in Hungarian 689
Renaissance and Reformation 689
The 17th century 689
  Effects of the Counter-Reformation
  Period of decline
The period of the Enlightenment 690
The 19th century 691
  Romanticism
  Writers of the late 19th century
The 20th century 691
  Early years
  The interwar period
  Writing after 1945
Bibliography 692

EARLIEST WRITINGS IN HUNGARIAN

The earliest known written traces of the Hungarian language are mostly proper names embedded in the Latin text of legal or ecclesiastical documents. The first continuous example of the Hungarian language is the *Halotti beszéd,* a short funeral oration written in about 1200, moving in its simplicity. Many translations from Latin were made in the 13th and 14th centuries, but the only one that has survived, and also the oldest extant poem written in Hungarian, is a free version of a poem by Godefroy de Breteuil. It is known as *Ómagyar Mária-siralom* (*c.* 1300; "Old Hungarian Lament of the Virgin Mary"). The 14th century also produced translations of the legends of St. Margaret and St. Francis of Assisi. The *Jókai codex,* which contains the St. Francis legend, was written in about 1440 and is the oldest extant Hungarian codex.

The 15th century saw the first translations from the Bible. The preachers Thomas and Valentine, followers of the Bohemian religious reformer Jan Hus, were responsible for this work, of which the prophetic books, the Psalms, and the Gospels have survived. A great part of the vocabulary, created for the purpose, is still in use. A number of sermons by the Franciscan Pelbárt Temesvári, originally written in Latin, have come down in Hungarian translations. Among other translations are the first Hungarian drama, *A három kórösztény leányról* (*c.* 1520; *On Three Christian Virgins*), translated from the Latin original of Hrosvitha; a translation of the legend of St. Catherine of Alexandria; and a translation of the Song of Solomon.

Hungarian literature was not entirely religious. The existence of a history of the Trojan War and of a Hungarian version of the Alexander romance can be inferred from their South Slavic translations. In the 14th century secular literature developed and literary forms were introduced from abroad.

RENAISSANCE AND REFORMATION

In 1367 the first Hungarian university was founded, at Pécs. About 100 years later King Matthias I Corvinus established the first Hungarian printing press. The King became known for his library and his patronage of foreign scholars; during his reign Latin literature in Hungary reached its peak in Janus Pannonius, who had been educated in Italy.

The 16th century brought changes. After the Battle of Mohács (1526) the Ottoman Turks occupied a large part of Hungary and the country was split into three. It is in the era of the Reformation that Hungarian national literature really began. Benedek Komjáti, Gábor Pesti, and János Sylvester, all of whom were disciples of the humanist Erasmus, translated parts of the Bible with philological accuracy. Pesti made a very readable translation of Aesop's fables and published a Latin–Hungarian dictionary. Sylvester published the first Hungarian grammar and, to show the adaptability of the vernacular to classical verse forms, wrote the first Hungarian poem in couplets. In 1541 he published a translation of the New Testament. The Hungarian Reformation

The second half of the 16th century saw the beginnings of Hungarian drama. *Comoedia Balassi Menyhárt árultátásáról* (1569; "Comedy on the Treachery of Menyhárt Balassi"), a satire by an unknown author, was among the most interesting literary achievements of the Reformation. Péter Bornemisza, the first important Protestant writer in Hungary, gave an entrancing view of Hungarian life, teeming with fresh observations, vivid descriptions, and original comments. His volume *Ördögi Kísértetekről* (1578; "On the Temptations of the Devil") offered an interesting consideration of moral and sexual problems in the 16th century. A poem of farewell, written on leaving the country, was one of the gems of early Hungarian poetry. His *Tragoedia magyar nyelven* (1558; "Tragedy in Hungarian"), though based on Sophocles' *Electra,* is a skillful adaptation of the play in the spirit of humanism.

Perhaps the greatest single literary achievement of the Hungarian Reformation was a translation of the Bible by Gáspár Károlyi and others (1590). The translation played a role in the development of Hungarian similar to that of the Authorized Version in English.

Up to the 16th century religious literature seems to have fared better than secular literature, in part because secular literature was not written down. The late 16th-century minstrels were more learned than their predecessors and in many cases were driven to their profession by difficult economic conditions. Perhaps the most important was Sebestyén Tinódi, by temperament more historian than poet. He described the wars against the Turks with remarkable accuracy, but his verse was monotonous. Péter Ilosvai Selymes was the author of a romance, *Az híres nevezetes Toldi Miklósnak jeles cselekedetiről* (1574; "The Story of the Remarkable Nicholas Toldi's Extraordinary and Brave Deeds"), which achieved great popularity in Hungary and served as a basis for a masterpiece by János Arany in the 19th century. This romance was the one original piece in the flow of the mere entertainment literature characteristic of the 16th century, the principal genre of which was the *széphistória* ("beautiful story"), adapted from western European originals. Perhaps the best was the *História egy Árgirus nevű királyfiról* (*c.* 1575; "The Story of the Prince Argirus") by Albert Gergei, from an Italian original but interwoven with Hungarian folklore.

A great poet emerged in Bálint Balassi (1554–94), who at first imitated Petrarch and various Neo-Latin poets but later displayed originality with a cycle of love poems of great beauty and emotional intensity. His songs of war, while reflecting the vicissitudes of fighting the Turk at the borders of the Christian world, celebrate nature and individual bravery in almost hymnlike tones. The poetry of his last years is imbued with a deep religious feeling; the imagery of the poems of his last creative period (the Coelie cycle of love songs as well as his religious verse) is coloured by Mannerism. Bálint Balassi

THE 17TH CENTURY

In the 17th century Hungary was still divided into three parts. The first, under Turkish rule, played no part in the

development of Hungarian literature. The second, under Habsburg rule, was open to Italian and German Roman Catholic influence; the third, Transylvania, was in close relationship with Dutch and English Protestant thought. The leading Protestant scholar and writer of the 17th century was János Apáczai Csere. His chief work was a Hungarian encyclopaedia in which he endeavoured to sum up the knowledge of his time. The work, published at Utrecht in 1653, marked a development in technical vocabulary.

Effects of the Counter-Reformation. By the end of the 16th century the Counter-Reformation was gaining momentum in western Hungary. A Jesuit cardinal, Péter Pázmány, a master of Hungarian prose, was outstanding as an orator and essayist. His writing was characterized by a vigorous and clear, though far from simple, style, use of popular expressions, and solid argument. His *Isteni igazságra vezérlő kalauz* (1613; "Guide to Divine Truth") was a refutation of non-Catholic religious doctrines and a masterpiece of Baroque prose.

Under the influence of the Jesuits, many Hungarian aristocrats returned to the Catholic faith and sent their sons to the Austrian Catholic universities and to Rome. The Italian Baroque, especially the influence of Tasso and Marino, is evident in the work of Miklós Zrínyi, a great Hungarian statesman and military commander. Most of his prose work was an exposition of political and strategic ideas. His greatest literary achievement was an epic, *Szigeti veszedelem* (1651; "The Peril of Sziget"), in 15 cantos, on the siege in 1566 of Szigetvár, which had been defended against the Turks by Zrínyi's great-grandfather. Though the influence of classical epics is clear, the work remains profoundly original and Hungarian. Another poet of this time, István Gyöngyösi, composed long narrative poems and also many epithalamia, or nuptial poems. He was inventive and handled rhyme with ease, and his work was read widely during the 17th and 18th centuries.

Period of decline. The period between 1700 and about 1770 was a time of decline and slow consolidation in Hungarian literature. Memoirs form what is best in the prose literature of the period. Among the most absorbing are the autobiography of the well-traveled Miklós Bethlen, a leading Transylvanian statesman, and the confessions and memoirs (written in Latin and French, respectively) of Ferenc Rákóczi II, exiled prince of Transylvania and leader of the anti-Habsburg insurrection of 1703–11. The *Törökországi levelek* ("Letters from Turkey"), written from 1717 to 1758 by Kelemen Mikes, a companion in exile of Rákóczi, were addressed to an imaginary aunt. In choosing the epistolary genre Mikes was inspired by French models, and his work stands out for its excellent style and wry humour. The *Metamorphosis Transylvaniae* of Péter Apor is a nostalgic reminiscence of Transylvania.

Memoirs and epistolary works

The poetry of this epoch has little to offer. The poems of László Amade were informed by a Rococo taste, both in form and in content; he mainly wrote poetry of gallantry and courtship. Another poet and translator, the versatile Ferenc Faludi, took his expressions from popular language and folk songs.

This period of literary decadence produced notable works only in the fields of history and history of literature. Some of them were written in Latin. Among historians, Mátyás Bél, György Pray, and István Katona are most important. The first historian of Hungarian literature, Dávid Czvittinger, composed the biographies of some 300 Hungarian writers. His work was continued and improved by Péter Bod, whose *Magyar Athénás* (1766; "Hungarian Athenaeum") deals with more than 500 Hungarian men of letters.

THE PERIOD OF THE ENLIGHTENMENT

The Hungarian Enlightenment was more receptive to French and English ideas than it was productive of original developments. The period between about 1772 and 1825, though immensely important in the development of the Hungarian spirit, produced few writers of the first rank.

With the publication in 1772 of the first literary work by György Bessenyei, a translation (from the French) of Alexander Pope's *Essay on Man*, the new era began. All of Bessenyei's works served a didactic purpose. His drama *Ágis tragédiája* (1772; "The Tragedy of Agis") was a somewhat creaking vehicle for his liberal ideas. His best work, *Tarimenes utazása* (1802–04; "Tarimenes' Journey"), the first real novel in Hungarian, was a bitter attack on everything that was opposed to the Enlightenment. With destructive irony, Bessenyei, an officer of the Hungarian Guards, examined the shortcomings of contemporary society. His personal influence induced several of his fellow officers—for example, Sándor Báróczi and Ábrahám Barcsay—to try to convey the ideas of the Enlightenment in Hungarian to a Hungarian public.

Spurred on by new ideas, but basically traditionalists, József Gvadányi and András Dugonics produced amusing works that were both of some literary merit and popular. Gvadányi's best work, *Egy falusi nótáriusnak budai utazása* (1790; "The Journey to Buda of a Village Notary"), is a defense of national and traditional values against encroaching foreign ideas. The novel *Etelka* (1788), by Dugonics, a sentimental love story in a historical setting, was the first Hungarian best-seller. Both Gvadányi and Dugonics used the language of the common people, and this was perhaps their greatest merit. Ádám Pálóczi Horváth left a collection of 450 poems, a treasure-house of authentic folk songs.

The end of the 18th century was a period of experiments with poetic language. The pioneers of the use of Greek and Latin metres in Hungarian verse (to which they are eminently suited) were followed by Benedek Virág, who imbued with poetic inspiration verse forms that for his predecessors were merely formal exercises. It fell to Dániel Berzsenyi, who published a single volume of poetry, in 1813, to show what use a great poet could make of classical metre. His ode "A Magyarokhoz" ("To the Hungarians"), his "Fohász" ("Prayer"), and his elegy "A közelítő tél" ("On the Nearing Winter") express the transitoriness of power and of friendship.

The ideas of the Enlightenment were not universally welcomed in Hungary. Traditionalist elements looked with distrust on any imported ideas, and the government was increasingly suspicious of a spirit of intellectual freedom, which it believed had led to the French Revolution and, in Hungary, to the Jacobin conspiracy of Martinovics, crushed in 1794. Several writers went to prison for harbouring radical views. The most talented among them, János Batsányi, secured his place in the history of Hungarian literature by his poem "A Franciaországi változásokra" (1789; "On the Changes in France"), a vigorous warning to all tyrants "to cast their watchful eyes on Paris."

Sentimentalism found its exponents in József Kármán and Gábor Dayka. Kármán's only work of importance, *Fanni hagyományai* (1794; "The Memoirs of Fanny"), is a novel of sentiment written in the form of letters and diary entries. Very much on the lines of Goethe's *Werther*, the work nevertheless marks an important step in the history of the Hungarian novel. Dayka, who was a poet, died too young for the full measure of his talent to be realized.

The first important lyric poet since Bálint Balassi was Mihály Csokonai Vitéz, who continued the purely Hungarian poetical tradition. His many songs to a woman named Lilla are a happy blend of playful grace and subtle thoughts. The influence of Rousseau is very noticeable in some of his longer philosophical poems. Alexander Pope's *Rape of the Lock* served as a source of inspiration for Csokonai's comic epic *Dorottya* (1804), but Csokonai's poem is original and his context very Hungarian. The language of the poem is vigorous, even vulgar, and the plot is full of hilariously comic situations.

The place of Sándor Kisfaludy in Hungarian literature is secured by his first work, *Kesergő szerelem* (1801; "Bitter Love"), a lyric cycle depending on a very thin narrative thread. Writing in an elaborate verse form of 12 lines, called the Himfy verse, which he devised himself, Kisfaludy displayed great ingenuity in finding new variations on the theme of unhappy love.

Ferenc Kazinczy, a mediocre poet but an influential man of letters, was the pivot of literary life for about 40 years. For his involvement in the conspiracy of Martinovics he paid with six years' imprisonment. He wanted a literature refined and limpid, neither baroque nor popular, and his

Ferenc Kazinczy

interest was focused on style. He became the head of the *neologi*, or linguistic innovators, who tried to renew and enrich the Hungarian language so that it could express the most elaborate concepts. The success of the language reform was due, to a large extent, to Kazinczy's efforts.

THE 19TH CENTURY

Romanticism. The literary revival initiated by Kazinczy continued after his death. The literary leadership of Hungary at the beginning of the 19th century was assumed by Károly Kisfaludy when, in 1822, he founded a literary magazine, *Aurora,* to which all the important writers of the period contributed. He was also the first representative of Romanticism and the first playwright to achieve popular success.

While Kisfaludy's tragedies were applauded all over the country, *Bánk bán* (the *bán* was a high Hungarian dignitary), one of Hungary's best tragedies, by József Katona, was published in 1821 but, for the time being, was overlooked. Set in the 13th century and written in vigorous prose, the play was a masterful combination of national and individual conflicts, and one of its characters, Tiborc, a poor peasant, has remained ever since a symbol of the oppressed.

Ferenc Kölcsey was a deputy in the Hungarian parliament and a brilliant orator; his literary criticism was of a high standard, though unduly severe. His later poems, which were grave but vigorous in thought and expression, often dealt with national problems; his impressive "Hymnusz" (1823) became the Hungarian national anthem. After Kisfaludy's death, Mihály Vörösmarty became a central figure in literary life, producing writings of value in every genre. In particular he succeeded with a long epic poem *Zalán futása* (1825; "The Flight of Zalán"), written in a Romantic vein but expressing a concern for contemporary problems. This concern is evident also in many of his best lyric poems and even in his symbolic fairy play *Csongor és Tünde* (1831; "Csongor and Tünde").

Hungarian novels

In Hungarian literature, poetry was far ahead of drama, and the novel seemed slow in taking root. Miklós Jósika, a disciple of Sir Walter Scott, was the first successful novelist. His first and best work, the historical novel *Abafi* (1836), marked a turning point for the genre. József Eötvös, who after the 1848 revolution became a political theorist, produced two of the best novels in 19th-century Hungarian literature—*A falu jegyzője* (1845; *The Village Notary*), a portrait of feudal life in his own time, and *Magyarország 1514-ben* (1847; "Hungary in 1514"), about György Dózsa's peasants' revolt. They possessed exceptional qualities of characterization, both of individuals and of periods, and were political manifestos in support of the oppressed and against the appalling injustices that led to revolutions—of which Eötvös nevertheless disapproved.

The folk song and ballad collections of János Erdélyi and János Kriza exerted an influence on the further development of Hungarian poetry. "Popular poetry is the only real poetry" was the opinion of Sándor Petőfi, one of the greatest Hungarian poets, whose best poems rank among the masterpieces of world literature. He was an innovator and made a break with conventional subjects and poetic language. His poems are striking in immediacy of perception and directness of language and cover a vast range of subjects. The fervour of his patriotic poems inspired the revolution of 1848. Petőfi's many songs are enchanting in their simplicity, and in this genre he remained unsurpassed.

János Arany shared Petőfi's conviction of the value of popular poetry, but his approach was different, for his subjects were often taken from history and showed deep understanding of the human mind. He had the assurance of one who knew that what he wrote was the language of the people, lifted to a degree never surpassed in Hungarian. His ballads, often romantic, had vigour, conciseness, and uncommon evocative power. His great narrative poems, the *Toldi* trilogy (1847–79) and *Buda halála* (1864; *The Death of King Buda*), reflected eternal human problems; Arany's philosophy appeared through his characters and not in lengthy digressions and was accompanied by subtle humour.

Writers of the late 19th century. The peaks of poetry reached by Petőfi and Arany remained inaccessible to other poets during the rest of the 19th century. Hungary, after being defeated in the war of independence of 1848–49, was ruled from Vienna until 1867. External political pressures on Austria and the willingness of Hungarian society to end passive resistance made possible the Settlement (or Compromise) of 1867, which created the Austro-Hungarian Monarchy. The post-1867 industrial boom and Hungary's fast technological and commercial development produced a mood of complacency that was first broken by László Arany (son of János), whose ironic novel in verse, *A délibábok hőse* (1873; "The Hero of the Mirages"), is representative of the mood of disillusionment. Another poet, János Vajda, bridged the gap between the romantic populism of Petőfi and fin-de-siècle decadence: a gloomy visionary, with equal propensity for self-pity and self-aggrandizement, he was nevertheless an important innovator in the field of metaphor and poetic imagery.

The national theatre

In 1837 a national theatre was established to produce works of merit, but, with few exceptions, the standard of plays was low. Ede Szigligeti, a prolific playwright, wrote entertaining comedies and created a special genre of plays, the *népszínmű,* that give an idealized picture of village life but also contain a measure of social criticism. Very different were the plays of Imre Madách, whose masterpiece *Az ember tragédiája* (1861; *The Tragedy of Man*) dealt with universal human problems. This poetic drama followed man's destiny from creation through stages of history into a future of a phalanstery (a Utopian commune) and the ultimate extinction of life. The play was first staged in 1883 and remains a favourite with the Hungarian public.

The first outstanding novelist, Zsigmond Kemény, displayed, in such novels as *Zord idő* (1862; "Grim Times"), *A rajongók* (1858–59; "The Fanatics"), and *Férj és nő* (1852; "Husband and Wife"), a masterly skill in psychological analysis. His characters' own deeds determined their gloomy ends. Analysis often took the place of action in Kemény's novels, which were therefore difficult to read and not popular. On the other hand, Mór Jókai was a popular Hungarian novelist, an exceptional storyteller able to evoke any epoch and any milieu. His characters were idealized, and his descriptions tended to be brilliant rather than accurate. Among his numerous works (he published more than 200 books in his lifetime) were historical novels on problems of contemporary society. *Az arany ember* (1873; "That Golden Man"; Eng. trans., *Timár's Two Worlds*) is one of his best novels. Kálmán Mikszáth was also popular: he recorded with keen observation and sly humour the shortcomings of society but, although a politician and a member of parliament, was little concerned with improvement. Though the principal works of Géza Gárdonyi were published early in the 20th century, they belonged to the 19th century. *Egri csillagok* (1901; "The Stars of Eger") and *A láthatatlan ember* (1902; "The Invisible Man"; Eng. trans., *Slave of the Huns*) were well constructed.

During the 19th century, literary life in Hungary became organized to some extent: the Kisfaludy and Petőfi societies, founded in 1836 and 1876, respectively, were particularly influential, though the authority of the Hungarian Academy, founded in 1825, remained unchallenged. The principal critic of the second half of the century was Pál Gyulai. Literary criticism and the history of Hungarian literature attracted some of the best minds, including Jenő Péterfy and Frigyes Riedl.

The last third of the 19th century in Hungary was an era of literary decline in which writers based their work on social and political ideals that were becoming sterile. The great majority of Hungarian writers came from the nobility and lived as part of the middle class; only at the end of the century did lower-middle-class writers come to the fore. The periodical *A hét* ("The Week"), founded in 1890 by József Kiss, became the organ of a number of gifted writers, including Zoltán Ambrus and Sándor Bródy.

THE 20TH CENTURY

Early years. The year 1906, when Endre Ady burst upon the literary scene with his *Új versek* ("New Poems"),

The influence of the periodical *Nyugat*

marked a turning point. In matters of style Ady was influenced by the French Symbolists, but in content he was concerned with radical political ideas. He rejuvenated the language of Hungarian poetry, introducing new themes and powerful new imagery. His rise was helped by the periodical *Nyugat* ("The West"), which was launched in 1908 under the editorship of Hugo Ignotus, Miksa Fenyő, and Ernő Osvát. Among poets associated with *Nyugat* were Mihály Babits, an excellent translator of foreign poetry who became editor in 1929; Dezső Kosztolányi, who wrote with empathy on childhood and death and whose novels and short stories established high standards in narrative prose; and Árpád Tóth and Gyula Juhász, who voiced the distress of the poor and the oppressed in society. A fifth poet, Milán Füst, wrote little, but the dramatic metaphors and sonorous language of the work he did produce made his a lasting influence. In addition to his poetry he wrote an outstanding novel, *A feleségem története* (1942; "The Story of My Wife").

The prose writers of *Nyugat* included Zsigmond Móricz, whose tales of provincial life portrayed peasants and gentry; Margit Kaffka, the first major woman writer in Hungary; and Gyula Krúdy, who created a nostalgic dreamworld with his stream-of-consciousness technique.

Writers not connected with *Nyugat* included the versatile Ferenc Molnár, who, after a promising start as a writer of fiction, began to write cleverly constructed social comedies. A conservative-nationalist group of writers was highly influential before 1918; its principal figure was Ferenc Herczeg, an author of novels and plays. During World War I and the years of revolution that followed, two authors emerged to challenge both the old establishment and *Nyugat*. These were Lajos Kassák, the first significant poet of the Hungarian avant-garde, who also wrote a remarkable autobiography depicting working-class life at the beginning of the century; and Dezső Szabó, whose large, uneven expressionistic novel *Az elsodort falu* (1919; "The Village That Was Swept Away") combined antiwar sentiment with a romantic cult of the peasantry. First embraced and then rejected by the post-1919 counterrevolution, Szabó is best remembered as a witty though venomous pamphleteer.

The interwar period. The interwar period saw a flowering of Hungarian letters. Although the influence of *Nyugat* diminished, neither the populist *Válasz* ("The Response") nor the left-wing *Szép szó* ("Fine Word") could quite supplant it. The leading poet of the 1920s was Lőrinc Szabó, a master of poetic technique and fine observation, whereas the 1930s were dominated by Attila József, whose experience of alienation and Socialist ideas were expressed in great poetic tableaux and in poems probing the subconscious, and by Gyula Illyés, who found inspiration in the life of the peasantry. The poetry of Miklós Radnóti reached a tragic climax in the serene and polished poems he wrote in the last years of his life.

In Hungary, as elsewhere, the novel became the principal form of literary expression. While Sándor Márai and Lajos Zilahy depicted the life of the bourgeoisie, János Kodolányi, László Németh, and Zsigmond Remenyik exposed the conflicts of the individual with society (often against a background of injustice and misery). Áron Tamási wrote beautifully stylized novels on the life of the Szeklers, an ethnic group of Transylvania. Tibor Déry, whose chief work was published only after 1945, wrote realistic novels and a challenging autobiography. The foremost essayist of the period was László Németh, whose *A minőség forradalma* (1940; "The Revolution of Quality") remained a seminal influence for many years to come. Other essayists and literary historians active in this era included a particularly brilliant writer, Antal Szerb, and Gábor Halász (both died in forced labour camps in 1945) and László Cs. Szabó.

Writing after 1945. The period since 1945, though officially designated one of "Socialist transformation," has seen but little change in writers' traditional orientations and preoccupations. During the first decade, particularly the years 1948–53, many writers were forced into silence by the regime's attempts to introduce Socialist Realism as the only correct style and creative method. After the failure of the 1956 uprising a number of writers were imprisoned, but by the mid-1960s most efforts to enforce ideological purity in the arts were abandoned. Since then there has been comparatively little official intervention in Hungarian literature and the margin of free experimentation has grown. This allowed writers such as Géza Ottlik, Miklós Mészöly, and István Örkény to publish work that showed ways in which the technique of modern fiction could be applied in Hungary. Among the best new authors were György Konrád and Péter Esterházy. Konrád's novels *A látogató* (1969; *The Case Worker*), *A városalapító* (1977; *The City Builder*), and the unofficially published *A cinkos* (1982; *The Loser*) achieved great impact with their dense, poetically structured style and analytical probing into the world of the social caseworker, the planner of new society, and the mental institution. Esterházy's most successful novel to date, *Termelési regény (kisssregény)* (1979; "Production Novel (a Ssshort [sic] Novel)"), is a grotesque and refreshingly irreverent survey of Hungarian life and society.

Among the adherents of realistic fiction, József Lengyel, who died in 1975, occupied a special place. In his stories (which could not be published until the loosening of restrictions in the early 1960s) he gave a moving testimony of human suffering in Soviet labour camps.

The best poetry was written by Sándor Weöres, whose poetic span ranges from Eastern philosophy to delightful children's verses, and János Pilinszky, an Existentialist Catholic whose most memorable poems deal with the experience of what he called the "universe of camps" produced by World War II. Other noteworthy poets included the urbane László Kálnoky and István Vas, and Ferenc Juhász and László Nagy, two poets of peasant origin whose work grew out of native tradition to express universal rites and myths of mankind such as marriage, the struggle among generations for power, and cosmic destruction.

Frontier changes since World War I have placed substantial Hungarian minorities in countries outside Hungary, especially in neighbouring Czechoslovakia, Yugoslavia, and Romania. In Romania, for example, where approximately 2,000,000 Hungarians live, the best known Hungarian writer is the playwright and novelist András Sütő. There also has been a large diaspora in the West, where, apart from Márai, the versatile modernist Győző Határ and the post-Romantic poet György Faludy have the largest following. The Munich-based cultural review *Uj látóhatár* ("New Horizon") has enjoyed the longest uninterrupted existence among Hungarian periodicals inside or outside of Hungary.

BIBLIOGRAPHY. ALBERT TEZLA, *An Introductory Bibliography to the Study of Hungarian Literature* (1964), and *Hungarian Authors: A Bibliographical Handbook* (1970), are annotated sources on items available from major U.S. and European book collections. ANTAL SZERB, *Magyar irodalomtörténet*, 6th ed. (1978), is an informative history, an earlier edition of which is available also in a German translation: *Ungarische Literaturgeschichte*, 2 vol. (1975). An exhaustive historical coverage is provided in ISTVÁN SÖTÉR (ed.), *A magyar irodalom története*, 6 vol. (1964–66). BÉLA POMOGÁTS, *Az újabb magyar irodalom, 1945–1981* (1982), is an informative critical study of 20th-century literature. In English, see D. MERVYN JONES, *Five Hungarian Writers* (1966), which gives a detailed analysis of the works of the writers Miklós Zrínyi, Kelemen Mikes, Mihály Vörösmarty, József Eötvös, and Sándor Petőfi. TIBOR KLANICZAY (ed.), *A History of Hungarian Literature* (1982; originally published in Hungarian, 1982), provides broad coverage. LÓRÁNT CZIGÁNY, *The Oxford History of Hungarian Literature from the Earliest Times to the Present* (1984), is a comprehensive chronologically arranged work, with emphasis on the 19th and 20th centuries.

(T.Kl./D.Si./G.G.)

Hungary

Located in the heart of Europe and occupying part of the Carpathian Basin, Hungary (in full, the Republic of Hungary; Hungarian: Magyar Köztársaság) has an area of 35,919 square miles (93,030 square kilometres). The nation has ethnic and linguistic roots that reach far back into the past. Its boundaries, however, have changed repeatedly over the centuries as events in Europe precipitated the reduction, expansion, and partition of Hungarian territory.

Modern Hungary shares a border to the north with Slovakia, to the northeast with Ukraine, to the east with Romania, to the south with Serbia and Montenegro (the Vojvodina region of Serbia) and Croatia, to the southwest with Slovenia, and to the west with Austria. Budapest, the capital city, which dominates much of national life, is situated on both banks of the Danube (Hungarian: Duna) River just downstream from the Danube Bend.

This article is divided into the following sections:

Physical and human geography 693
- The land 693
  - Relief
  - Drainage and soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 696
  - Ethnic and religious structure
  - Demographic trends
- The economy 696
  - Overview
  - Resources
  - Agriculture
  - Industry
  - Finance
  - Trade
  - Transportation
- Administration and social conditions 698
  - Government
  - Judiciary
  - Armed forces
  - Education
  - Health and welfare
  - Housing
- Cultural life 698
  - Daily life
  - Arts
  - Cultural institutions
  - Recreation
  - Press and broadcasting

History 700
- The kingdom to 1526 700
  - The Árpáds
  - Hungary under foreign kings
- The period of partition 704
  - Royal Hungary and the rise of Transylvania
  - War and liberation
- Habsburg rule, 1699–1918 705
  - Habsburg rule to 1867
  - The Dual Monarchy, 1867–1918
- Revolution, counterrevolution, and the Regency, 1918–45 708
  - The Regency, 1920–45
  - Financial crisis: the rise of right radicalism
  - War and renewed defeat
- Hungary since 1945 710
  - The communist regime
  - The revolution of 1956
  - The Kádár regime
  - Reforms of the late 1980s
  - Stabilizing democracy and the market economy

Bibliography 714

Physical and human geography

THE LAND

The Great and Little Alfölds

Relief. Dominating the relief are the great lowland expanses that make up the core of Hungary. The Little Alfold (Little Hungarian Plain, or Kis Alföld) lies in the northwest, fringed on the west by the easternmost extension of the sub-Alps along the border with Austria and bounded on the north by the Danube. The Little Alfold is separated from the Great Alfold (Great Hungarian Plain, or Nagy Magyar Alföld) by a low mountain system extending across the country from southwest to northeast for a distance of 250 miles (400 kilometres). This system, the backbone of the country, is made up of Transdanubia (Dunántúl) and the Northern Mountains, separated by the Visegrád Gorge of the Danube. The former is dominated by the Bakony Mountains, with dolomite and limestone plateaus at elevations between 1,300 and 2,300 feet (400 and 700 metres) above sea level interspersed with volcanic peaks; the latter, which consist of volcanic rocks, comprise the Mátra Mountains in the north, reaching a height of 3,327 feet (1,014 metres) at Mount Kékes, the nation's highest peak. Regions of hills reaching elevations of 800 to 1,000 feet lie on either side of the mountain backbone, while to the south and west of Lake Balaton is an upland region of more subdued, loess-covered topography. The Great Alfold covers most of central and southeastern Hungary. Like its northwestern counterpart, it is a basinlike structure filled with fluvial and windblown deposits. Four types of surface may be distinguished: floodplains, composed of river alluvium; alluvial fans, wedge-shaped features deposited at the breaks of slopes where rivers emerge from the mountain rim; alluvial fans overlain by sand dunes; and plains buried under loess, deposits of windblown material derived from the continental interior. These lowlands range in height from about 260 to 660 feet above sea level, with the lowest point at 256 feet, on the southern edge of Szeged, along the Tisza River.

Drainage and soils. Hungary lies within the drainage basin of the Danube. The Danube and two of its tributaries, the Rába and the Drava, are of Alpine origin, while the Tisza and its tributaries, which drain much of eastern Hungary, rise in the Carpathian Mountains to the east.

The Danube floods regularly twice a year, first in early spring and again in early summer. During these phases discharge is up to 10 times greater than that recorded during the low-water periods of autumn and winter. The Tisza forms a floodplain as it flows through Hungary, and large meanders and oxbow lakes marking former channels are typical features. Devastating floods have occurred on the Danube, the Tisza, and their tributaries. About 2,500 miles of levees have been built to protect against floods. The relatively dry climate of the central and eastern areas of the Great Alfold has necessitated the construction of large-scale irrigation systems, mostly along the Tisza River.

There are few lakes in Hungary, and most are small. Lake Balaton, however, is the largest freshwater lake in central Europe, at 231 square miles (598 square kilometres). Lake Fertő (Neusiedler Lake) lies on the Austrian border, and Lake Velence lies southeast of Budapest.

Gray-brown podzolic (leached) and brown forest soils predominate in the forest zones, while rich black-earth, or chernozem, soil has developed under the forest steppe. Sand dunes and dispersed alkali soils are also characteristic.

Climate. Because of its situation within the Carpathian Basin, Hungary has a moderately dry continental climate.



Cities over 1,000,000
Cities 100,000 to 1,000,000
Cities 30,000 to 100,000
Cities under 30,000

National capitals
County capitals
TOLNYA County names
International boundaries
County boundaries

Canals
Swamps and marshes
National parks
Spot elevations in metres (1 m = 3.28 ft)

Scale 1:3,361,000
1 inch equals approx. 53 miles
0 10 20 30 40 50 mi
0 20 40 60 80 km
Conic Projection

MAP INDEX

Political subdivisions

Bács-Kiskun . . . . . 46 30 N 19 25 E
Baranya . . . . . 46 05 N 18 15 E
Békés . . . . . 46 45 N 21 00 E
Borsod-Abaúj-Zemplén . . . . . 48 15 N 21 00 E
Budapest . . . . . 47 30 N 19 05 E
Csongrád . . . . . 46 25 N 20 15 E
Fejér . . . . . 47 10 N 18 35 E
Győr-Moson-Sopron . . . . . 47 40 N 17 15 E
Hajdú-Bihar . . . . . 47 25 N 21 30 E
Heves . . . . . 47 50 N 20 15 E
Jász-Nagykun-Szolnok . . . . . 47 15 N 20 30 E
Komárom-Esztergom . . . . . 47 37 N 18 20 E
Nógrád . . . . . 48 00 N 19 35 E
Pest . . . . . 47 25 N 19 20 E
Somogy . . . . . 46 25 N 17 35 E
Szabolcs-Szatmár-Bereg . . . . . 48 00 N 22 10 E
Tolna . . . . . 46 30 N 18 35 E
Vas . . . . . 47 10 N 16 45 E
Veszprém . . . . . 47 10 N 17 40 E
Zala . . . . . 46 40 N 16 50 E

Cities and towns

Abony . . . . . 47 11 N 20 00 E
Ajka . . . . . 47 06 N 17 34 E
Badacsonytomaj . . . . . 46 48 N 17 31 E
Baja . . . . . 46 11 N 18 58 E
Balassagyarmat . . . . . 48 05 N 19 18 E
Balatonfüred . . . . . 46 57 N 17 53 E
Balmazújváros . . . . . 47 37 N 21 21 E
Barcs . . . . . 45 58 N 17 28 E
Bátonyterenye . . . . . 47 58 N 19 50 E
Békés . . . . . 46 46 N 21 08 E
Békéscsaba . . . . . 46 41 N 21 06 E
Berettyóújfalu . . . . . 47 13 N 21 33 E
Bicske . . . . . 47 29 N 18 38 E
Bonyhád . . . . . 46 18 N 18 32 E
Budakeszi . . . . . 47 31 N 18 56 E
Budaörs . . . . . 47 27 N 18 58 E
Budapest . . . . . 47 30 N 19 05 E
Cegléd . . . . . 47 10 N 19 48 E
Celldömölk . . . . . 47 15 N 17 09 E
Csongrád . . . . . 46 42 N 20 09 E
Csorna . . . . . 47 37 N 17 15 E
Dabas . . . . . 47 11 N 19 19 E
Debrecen . . . . . 47 32 N 21 38 E
Dombóvár . . . . . 46 23 N 18 07 E
Dorog . . . . . 47 43 N 18 44 E
Dunaharaszti . . . . . 47 21 N 19 05 E
Dunakeszi . . . . . 47 38 N 19 08 E
Dunaújváros (Sztálinváros) . . . . . 46 59 N 18 56 E
Eger . . . . . 47 54 N 20 23 E
Esztergom . . . . . 47 48 N 18 45 E
Fertőd (Eszterháza) . . . . . 47 37 N 16 52 E
Fót . . . . . 47 37 N 19 12 E
Gödöllő . . . . . 47 42 N 19 08 E
Gyál . . . . . 47 25 N 19 22 E
Gyomaendrőd . . . . . 47 23 N 19 14 E
Gyömrő . . . . . 47 25 N 19 24 E
Gyöngyös . . . . . 47 46 N 20 50 E
Győr . . . . . 47 41 N 17 38 E
Gyula . . . . . 46 39 N 21 17 E
Hajdúböszörmény . . . . . 47 40 N 21 31 E
Hajdúnánás . . . . . 47 51 N 21 26 E
Hajdúszoboszló . . . . . 47 27 N 21 24 E
Hatvan . . . . . 47 40 N 19 41 E
Heves . . . . . 47 36 N 20 17 E
Hévíz . . . . . 46 47 N 17 11 E
Hódmezővásárhely . . . . . 46 25 N 20 20 E
Jánoshalma . . . . . 46 18 N 19 20 E
Jászapáti . . . . . 47 31 N 20 09 E
Jászberény . . . . . 47 30 N 19 55 E
Kalocsa . . . . . 46 32 N 19 00 E
Kaposvár . . . . . 46 22 N 17 48 E
Kapuvár . . . . . 47 36 N 17 02 E
Karcag . . . . . 47 19 N 20 56 E
Kazincbarcika . . . . . 48 15 N 20 38 E
Kecskemét . . . . . 46 54 N 19 42 E
Kerepestarcsa . . . . . 47 32 N 19 16 E
Keszthely . . . . . 46 46 N 17 15 E
Kiskőrös . . . . . 46 37 N 19 18 E
Kiskunfélegyháza . . . . . 46 43 N 19 51 E
Kiskunhalas . . . . . 46 26 N 19 30 E
Kisújszállás . . . . . 47 13 N 20 46 E
Kisvárda . . . . . 48 13 N 22 05 E
Komárom . . . . . 47 44 N 18 07 E
Komló . . . . . 46 12 N 18 16 E
Körmend . . . . . 47 01 N 16 36 E
Kőszeg . . . . . 47 23 N 16 33 E
Kőváguszőlős . . . . . 46 05 N 18 08 E
Lenti . . . . . 46 37 N 16 33 E
Makó . . . . . 46 13 N 20 29 E
Marcali . . . . . 46 35 N 17 25 E
Mátészalka . . . . . 47 57 N 22 20 E
Mezőberény . . . . . 46 49 N 21 02 E
Mezőkövesd . . . . . 47 49 N 20 35 E
Mezőtúr . . . . . 47 00 N 20 38 E
Miskolc . . . . . 48 06 N 20 47 E
Mohács . . . . . 45 59 N 18 42 E
Monor . . . . . 47 21 N 19 27 E
Mór . . . . . 47 23 N 18 12 E
Mosonmagyaróvár . . . . . 47 52 N 17 17 E
Nagyatád . . . . . 46 13 N 17 22 E
Nagykanizsa . . . . . 46 27 N 16 59 E
Nagykáta . . . . . 47 25 N 19 45 E
Nagykőrös . . . . . 47 02 N 19 47 E
Nyírbátor . . . . . 47 50 N 22 08 E
Nyíregyháza . . . . . 47 57 N 21 43 E
Orosháza . . . . . 46 34 N 20 40 E
Oroszlány . . . . . 47 29 N 18 19 E
Ózd . . . . . 48 13 N 20 18 E
Paka . . . . . 46 38 N 18 52 E
Pápa . . . . . 47 20 N 17 28 E
Pécs . . . . . 46 05 N 18 14 E
Pomáz . . . . . 47 39 N 19 02 E
Püspökladány . . . . . 47 19 N 21 07 E
Rudabánya . . . . . 48 23 N 20 38 E
Sajószentpéter . . . . . 48 13 N 20 43 E
Salgótarján . . . . . 48 07 N 19 49 E
Sárbogárd . . . . . 46 53 N 18 38 E
Sarkad . . . . . 46 45 N 21 23 E
Sárospatak . . . . . 48 19 N 21 35 E
Sárvár . . . . . 47 15 N 16 56 E
Sátoraljaújhely . . . . . 48 24 N 21 40 E
Siklós . . . . . 45 51 N 18 18 E
Siófok . . . . . 46 54 N 18 03 E
Sopron . . . . . 47 41 N 16 36 E
Szarvas . . . . . 46 52 N 20 33 E
Szeged . . . . . 46 15 N 20 10 E
Szeghalom . . . . . 47 02 N 21 10 E
Székesfehérvár . . . . . 47 12 N 18 25 E
Szekszárd . . . . . 46 21 N 18 43 E
Szentes . . . . . 46 39 N 20 16 E
Szigetszentmiklós . . . . . 47 21 N 19 03 E
Szigetvár . . . . . 46 03 N 17 48 E
Szolnok . . . . . 47 11 N 20 12 E
Szombathely . . . . . 47 14 N 16 37 E
Sztálinváros, see Dunaújváros
Tamási . . . . . 46 38 N 18 17 E
Tapolca . . . . . 46 53 N 17 26 E
Tata . . . . . 47 39 N 18 19 E
Tatabánya . . . . . 47 34 N 18 25 E
Téglás . . . . . 47 41 N 21 40 E
Tiszafüredvár . . . . . 46 59 N 20 15 E
Tiszafüred . . . . . 47 37 N 20 46 E
Tiszaújváros . . . . . 47 56 N 21 05 E
Tiszavasvári . . . . . 47 58 N 21 21 E
Tokaj . . . . . 48 07 N 21 25 E
Törökszentmiklós . . . . . 47 11 N 20 25 E
Türkeve . . . . . 47 06 N 20 45 E
Újfehértó . . . . . 47 48 N 21 41 E
Vác . . . . . 47 47 N 19 08 E

Várpalota ..... 47 12 N 18 08 E
Vecsés ..... 47 24 N 19 17 E
Veszprém ..... 47 06 N 17 55 E
Visegrád ..... 47 47 N 18 59 E
Zalaegerszeg ..... 46 50 N 16 51 E

Physical features and points of interest

Bakony Mountains ..... 47 15 N 17 50 E
Balaton, Lake ..... 46 50 N 17 45 E
Baranya, see Mecsek
Bükk Mountains ... 48 05 N 20 30 E
Bükk National Park ..... 48 06 N 20 30 E
Csóványos, Mount ..... 47 57 N 18 57 E
Danube (Duna), *river* ..... 47 00 N 19 00 E
Drava (Dráva), *river* ..... 45 33 N 18 55 E
Dunántúl, see Transdanubia
Dunavölgyi Canal ..... 46 12 N 18 56 E
Kőris, Mount ..... 47 18 N 17 45 E
Eger, *river* ..... 47 38 N 20 39 E
Fertő (Neusiedler), Lake ..... 47 50 N 16 45 E
Great Alfold (Great Hungarian Plain, or Nagy Magyar Alföld), *lowland* ..... 47 00 N 20 00 E
Hajdúság, *region* .. 47 35 N 21 30 E
Hortobágy, *region* . 47 35 N 21 05 E
Hortobágy National Park ..... 47 34 N 21 09 E
Istállóskő, Mount .. 48 04 N 20 26 E
Kapos, *river* ..... 46 44 N 18 29 E
Kékes, Mount .... 47 52 N 20 01 E
Keleti Main Canal ..... 48 01 N 21 20 E
Kis Alföld, see Little Alfold
Kiskörei Reservoir ..... 47 30 N 20 40 E
Kiskunság, *region* .. 46 35 N 19 15 E
Kiskunság National Park .... 46 40 N 19 30 E
Kőris, Mount ..... 47 18 N 17 45 E
Körös, *river* ..... 46 43 N 20 12 E
Little Alfold (Little Hungarian Plain, or Kis Alföld), *lowland* ..... 47 30 N 17 00 E
Marcal, *river* ..... 47 38 N 17 32 E
Maros (Mureşul), *river* ..... 46 15 N 20 12 E
Mátra Mountains .... 47 53 N 19 57 E
Mecsek (Baranya) Mountains ..... 46 10 N 18 18 E
Mura (Mur), *river* ... 46 18 N 16 55 E
Nagy Magyar Alföld, see Great Alfold
Neusiedler, see Fertő
Nógrádi, *river* ..... 47 40 N 19 41 E
Northern Mountains, *region* ..... 48 00 N 20 30 E
Nyírség, *region* .... 47 50 N 21 55 E
Nyugati Main Canal ..... 47 57 N 21 19 E
Rába, *river* ..... 47 41 N 17 38 E
Sió, *river* ..... 46 22 N 18 48 E
Tisza (Tisa), *river* ..... 45 15 N 20 17 E
Tiszántúl, see Transtisza
Transdanubia (Dunántúl), *region* ..... 47 00 N 18 00 E
Transtisza (Tiszántúl), *region* ..... 47 00 N 21 00 E
Velence (Velencei), Lake .. 47 13 N 18 36 E

Temperatures

The mean annual temperature is about 50° F (10° C). Average temperatures range from 32° to 25° F (0° to −4° C) in January to 64° to 73° F (18° to 23° C) in July. Recorded temperature extremes are 109° F (43° C) in summer and −29° F (−34° C) in winter. In the lowlands, precipitation generally ranges from 20 to 24 inches (500 to 600 millimetres), rising to 24 to 31 inches over higher elevations. The central and eastern areas of the Great Alfold are the driest and the southwestern uplands the wettest part of the country. As much as two-thirds of annual precipitation falls during the growing season.

Plant and animal life. Human activities over the ages have largely destroyed the natural vegetation of Hungary. More than half of the land is regularly cultivated, and a sixth is used for nonagricultural purposes. The remainder comprises meadows and rough pasture and forest and woodland. No part of the country is of sufficient elevation to support natural coniferous forest, beech being the climax species at the highest elevations and oak woodland alternating with scrubby grassland being the climax at lower elevations in the upland regions.

Deer and wild pigs are abundant in the forests at higher elevations, while rodents, hares, partridge, and pheasant inhabit the lowlands. The once numerous varieties of marsh waterfowl survive only in nature reserves. There are diverse species of freshwater fish, including pike, bream, and pike perch. Significant water and air pollution occurs in some of the industrial regions of the country.

Settlement patterns. *Traditional regions.* The Great Alfold is the largest region of the country. It is divided into two parts: Kiskunság, the area lying between the Danube and Tisza rivers, and Transtisza (Tiszántúl), the region east of the Tisza. The former consists primarily of a mosaic of small landscape elements—sand dunes, loess plains, and floodplains. Kecskemét is the market centre for the region, which is also noted for its isolated farmsteads, known as *tanyák*. Several interesting groups live there, including the people of Kalocsa and the Matyó, who occupy the northern part of the plain around Mezőkövesd and are noted for hand embroidery and multicoloured costumes. In the generally homogeneous flat plain of the Transtisza region, only the Nyírség region in the northeast presents any form of topographical contrast. Closely connected with this latter area are the Hajdúság and the Hortobágy regions, and all three areas look to Debrecen, the largest city of the plain. The Hortobágy presents an interesting survival of the steppe life of earlier times, since the original Hungarian cattle, horse, and sheep breeds have been preserved there as part of the national heritage.

The Little Alfold, the second major natural region, is situated in the northwest and is traversed by the Danube and Rába rivers and their tributaries. It is more favourably endowed with natural resources than is the Great Alfold; both agriculture and industry are more advanced there. Győr is the major city of the region.

Transdanubia

The third major region, Transdanubia, embraces all the country west of the Danube exclusive of the Little Alfold. It is a rolling upland broken by the Bakony and Mecsek ridges. Lake Balaton is a leading resort area. To the south of the lake are the hills of Somogy, Tolna, and Baranya counties, where Pécs, a mining and industrial city, is the economic and cultural centre. Also found in Transdanubia are the Bakony Mountains, whose isolation, densely forested ridges, small closed basins, and medieval fortresses and monasteries have protected the local inhabitants over the course of many stormy centuries. Although modern industrial towns drawing on the bauxite, manganese, and brown coal resources of the area have sprung up, the cultural centre of Transdanubia is the historic city of Veszprém. In the southern part of the region, north and west of Lake Balaton, are health resorts and centres of wine production, notably Keszthely, Hévíz, Badacsony, and Balatonfüred.

The Northern Mountains, the fourth major geographic region of the country, contain two important industrial areas, the Nógrád and Borsod basins. Agriculture is important, especially viticulture, including the well-known Tokaj (Tokay) and Eger vineyards. Tourism is well-developed, and numerous spas and recreation centres are located there. Miskolc is the main economic centre for the region.

Urban settlement. Nearly two-thirds of the population is urban, but the majority of towns in Hungary have populations of less than 40,000. They were functionally, until the late 20th century, vastly overgrown villages rather than towns. About one-third of the urban population lives within the Budapest metropolitan area.

Urban Hungary is dominated by Budapest, which is several times the size of other major cities. It has the largest industrial workforce in the country. The major provincial centres are Miskolc, Debrecen, Szeged, Pécs, and Győr, each of which has a population exceeding 100,000; an economic, cultural, and administrative hinterland that reaches deep into the surrounding countryside; and an expanding industrial capacity. Below the provincial centres in the hierarchy are the traditional market towns, such as Kecskemét, Székesfehérvár, Nyíregyháza, Szombathely, and Szolnok, with new suburbs added to medieval or Baroque town centres. Also worthy of note are the predominantly industrial towns located close to the mineral resources of the Northern Mountains, which, from small beginnings in the late 19th century, have developed into major industrial centres. They include Tatabánya, Salgótarján, and Ózd. In addition, a number of industrial towns were created in the late 20th century on green-field sites as part of deliberate planning policy. These include the metallurgical centre of Dunaújváros on the Danube and the chemical centre of Kazincbarcika in eastern Hungary.

Rural population distribution

Rural settlement. The distribution of rural population varies widely from one part of the country to another. For historical reasons connected with resettlement following the Turkish occupation in the 16th century, the compact villages of the Great Alfold are small in number but large in size. By comparison, rural settlement in Transdanubia and in the Northern Mountains takes the form of many small nucleated and linear villages. The *tanyák* tend to be concentrated in the Great Alfold.

THE PEOPLE

Ethnic and religious structure. From its inception in the 11th century, the Kingdom of Hungary was a multiethnic country. Major territorial changes made Hungary ethnically homogenous after World War I, however, and more than nine-tenths of the population is now ethnically Hungarian and speaks Hungarian (Magyar) as its mother tongue. The Hungarian language is classified as a member of the Ugric branch of the Uralic languages; as such it is most closely related to the Ob-Ugric languages, Khanty and Mansi, which are spoken east of the Ural Mountains. It is also related, though more distantly, to Finnish and Estonian, each of which is (like Hungarian) a national language; to the Sami languages of far northern Scandinavia; and, more distantly still, to the Samoyedic languages of Siberia. Ethnic Hungarians are a mix of the Finno-Ugric Magyars and various assimilated Slavic, Turkish, and Germanic peoples. About 3 percent of the population is Gypsy (Rom), and nearly another 5 percent is made up of Slovaks, Romanians, Croats, Germans, and others.

Language

More than two-thirds of the people are Roman Catholic, most of them living in the western and northern parts of the country. About one-fifth of the population is Calvinist (concentrated in eastern Hungary), and relatively smaller groups belong to various Christian denominations (Lutherans, Greek Catholics, Eastern Orthodox, and Unitarians). The Jewish community, which constituted 5 percent of the population before World War II, has represented less than 1 percent of the population since the Holocaust.

Demographic trends. Owing to major changes in Hungary's borders following World War I, the republic's population decreased to less than eight million. Since then population growth has been rather slow, as both birth and death rates generally declined in the post-World War II years, though the latter returned to prewar levels by the mid-1980s. Life expectancy for women increased consistently from the 1930s; that for men also increased until the 1970s, when the trend reversed, and women now outlive men by about 10 years.

The population reached its peak in 1980 at some 10.7 million, but, because death rates exceeded birth rates, it decreased thereafter. Demographic trends were influenced by the modernization process and by a rate of suicide that was among the highest in the world.

Roughly three million Hungarians live in the neighbouring countries of Romania, Slovakia, Croatia, and Serbia. As a consequence of a net overseas emigration of 1.3 million people before World War I and a continuous, though much smaller, emigration related to major political upheavals in 1918–19, the 1930s, 1944, and 1956, large Hungarian communities live in North America and western Europe. After the collapse of communism and the splintering of Yugoslavia, roughly 100,000 refugees immigrated to Hungary from Romania and the former Yugoslav federation. Half of them were ethnic Hungarians.

Emigration

THE ECONOMY

Overview. Hungary remained mostly agrarian until World War II. Beginning in 1948, a forced industrialization policy based on the Soviet pattern changed the economic character of the country. A centrally planned economy was introduced and millions of new jobs were created in industry (notably for nonworking women) and, later, in services. This was accomplished largely through a policy of forced accumulation: keeping wages low and the prices of consumer goods (as opposed to staples) high made it possible for more people to be employed, and, because consumer goods were beyond their means, most Hungarians put more of their earnings in savings, which became available for use by the government. In the process the proportion of the population employed in agriculture declined from more than half to one-eighth by the 1990s, while the industrial workforce grew to nearly a third of the economically active population by the late 1980s.

Although Soviet-type economic modernization generated rapid growth, it was based on an early 20th-century structural pattern and on outdated technology. The heavy industries of iron, steel, and engineering were given the highest priority, while modern infrastructure, services, and communication were neglected. Moreover, the lack of entrepreneurial interest and market incentives prevented the development of new technologies and high-tech industries, as did Western restrictions (the Coordinating Committee for Multilateral Export Controls) on the export of modern technology to the Soviet bloc.

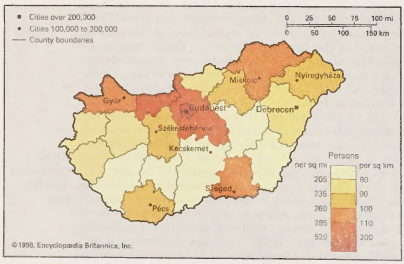
In the late 1960s a mixed economy was introduced in Hungary. Market prices and incentives gradually gained ground, and a partial privatization program was initiated. By the end of the 1980s one-third of the gross domestic product (GDP)—nearly three-fifths of services and more than three-fourths of construction—was being generated by private business. The Hungarian economy, however, failed to meet the challenge of the world economic crisis after 1973. The dramatic price increases for oil and modern technology created a large external trade deficit, which led to increasing foreign indebtedness. Growth slowed down and inflation rose, leading to a period of stagflation.

After 1989 Hungary's emerging market and parliamentary systems inherited a crisis-ridden economy with an enormous external debt and noncompetitive export sectors. Hungary turned to the world market and restructured its foreign trade, but market competition, together with a sudden and radical opening of the country and the abolition of state subsidies, led to further economic decline. Agriculture was drastically affected and declined by half. A large portion of the iron, steel, and engineering sectors, especially in northeastern Hungary, collapsed. Industrial output and GDP decreased by 30 percent and 25 percent, respectively. Unemployment, previously nonexistent, rose to 14 percent in the early 1990s but declined after 1994.

Period of transition

By the mid-1990s the economy was again growing, but only moderately. Inflation peaked in 1991 and remained high, at more than 20 percent annually, until the second half of the decade. As a consequence of unavoidable austerity measures that included the elimination of many welfare institutions, most of the population lost its previous security; the number of people living below the subsistence level increased from 10 to about 30 percent of the population between 1988 and 1995. Adjustment to the world economy, however, is evident. Major multinational companies made investments in Hungary that represented more than half of the entire international capital investment in central and eastern Europe in the first half of the 1990s. Modernization of telecommunications also began, and new modern industries (*e.g.*, automobile manufacturing) emerged. Significantly, some 750,000 small-scale, mostly family-owned enterprises were established, while state ownership of businesses declined to roughly one-fifth. Another important contributor to economic growth has been a flourishing tourist industry, as the number of foreign tourists reached more than 20 million by the mid-1990s.

Resources. The most important natural endowments of Hungary, particularly in its western and central areas, are its fertile soil and abundant water resources (notably Lake Balaton, a major asset for tourism). Half of the country's land is arable, and less than one-fifth is covered by woods. The climate is also favourable for agriculture.



Population density of Hungary.

Fossil fuel resources are relatively modest. High-quality anthracite (hard coal) is extracted only at Komló, and lignite (brown coal) is mined in the Northern Mountains (notably at Ózd) and in Transdanubia (at Tatabánya). Coal once satisfied half of Hungary's energy requirements; it now represents less than a third of energy production.

Oil and natural gas were discovered in the late 1930s in Transdanubia and during the postwar decades at several localities in the Great Alfold. Their share of energy production increased from one-third to one-half between 1970 and 1990; however, Hungary is able to meet only a fraction of its oil requirements with domestic resources.

The country's only significant mineral resources are bauxite—of which Hungary has some of the richest deposits in Europe—manganese, in the Bakony Mountains, and the undeveloped copper and zinc resources at Recsk. Extraction of various metal-bearing ores increased significantly in postwar Hungary, but iron ore is no longer mined. Other minerals that are found include mercury, lead, uranium, perlite, molybdenum, diatomite, kaolin, bentonite, zeolite, and dolomite.

Agriculture. Agriculture's role in the Hungarian economy declined steadily following World War II, dropping from half of the GDP in the immediate postwar period to less than one-tenth by the mid-1990s. Nevertheless, agriculture remains a leading economic sector. Two-thirds of the total area of the country is under cultivation, and Hungary is virtually self-sufficient in food production. Agriculture accounted for nearly one-fourth of Hungarian exports before the economic transition of the 1990s, during which animal stocks decreased by one-third and agricultural output and exports declined by half.

After the initial period of collectivization (1948–61), Hungarian cooperatives incorporated private farming. Private plots comprised roughly one-eighth of a cooperative's land and produced one-third of the country's agricultural output. One-fifth of Hungarian farmland belonged to state farms. Since 1990 the land has been reprivatized. Although farmers are owners of their land, many among the mostly elderly agricultural population remain in reorganized collective farms, producing one-third of agricultural output. Private farms, however, are the norm and produce more than half of the output.

Major crops

Cereals, primarily wheat and corn (maize), are the republic's most important crops. Other major crops are sugar beets, sunflower seeds, potatoes, and fruit. Viticulture, found in the Northern Mountain region, is also significant. Cattle, sheep, pigs, and poultry are raised in Hungary, but, in response to the government's efforts to combat overproduction of animal products, substantial reductions in livestock occurred in the 1990s.

Industry. As a result of the policy of forced industrialization under communism, industry experienced an exceptionally high growth rate until the late 1980s, by which time it constituted two-fifths of GDP. Mining and metallurgy, as well as the chemical and engineering industries, grew in leaps and bounds as the preferred sectors of Hungary's planned economy. Indeed, half of industrial output was produced by these three sectors. Without the implementation of modern technology and an efficient structure, however, Hungarian industry was not prepared to compete in the global economy after the collapse of state socialism. During the first half of the 1990s, industrial employment dropped to one-fourth of the economically active population. Total output declined by nearly one-third, with output in the mining, metallurgy, and engineering industries decreasing by half.

As industry and the Hungarian economy in general underwent restructuring and modernization during the early 1990s (including the implementation of privatization and the improvement of the quality of goods and services), some industries adapted more successfully to new conditions. Among the industries that regressed least and showed the first signs of growth were the food, tobacco, and wood and paper industries. Of Hungary's traditionally strong sectors, the chemical industry showed the greatest resilience, demonstrating growth again by the mid-1990s after experiencing a large drop in production (25 percent) early in the decade. Partly through foreign investment, the machine industry also showed signs of improvement by the mid-1990s. A number of newer industries, including the production and repair of telecommunication equipment and the automobile industry, showed significant growth.

Energy

Between 1950 and 1990, electric-power consumption in Hungary increased 10-fold, and by the 1990s more than one-third of industrial output was produced by the energy sector. Because nearly two-thirds of energy consumption is derived from thermal plants burning hydrocarbons (most of which must be imported), nuclear power is growing in importance, accounting for nearly one-fourth of energy production by the early 1990s. As the republic sought to reduce its dependency on foreign oil, a number of power plants were converted from oil to coal. A small percentage of power generation consists of hydroelectricity.

Finance. Under the Soviet-style, single-tier banking system, the National Bank both issued money and monopolized the financing of the entire Hungarian economy. Beginning in 1987, Hungary moved toward a market-oriented, two-tier system in which the National Bank remained the bank of issue but in which commercial banks were established. Foreign investment was permitted, and "consortium" (partly foreign-owned) banks were formed. In 1989 a stock exchange was established. In the 1990s, after the collapse of state socialism, the reform process continued with the founding of private banks, the sale of shares in state-owned banks (though most banks remained state-owned), and the enactment of a law that guaranteed the independence of the National Bank. The currency (forint) also became entirely convertible for business.

Trade. Hungary was a charter member of the Council for Mutual Economic Assistance (Comecon; 1949–91). Under its aegis trade was conducted between the countries of the Soviet bloc on the basis of specialized production, fixed prices, and barter. The Soviet Union was Hungary's most important trading partner, but in the late 1980s and early '90s, as Hungary became increasingly involved in the global market, less than half of the republic's trade remained with Comecon. Unprepared for the competitiveness of global market forces, Hungary accrued a large trade deficit that was covered by foreign loans. In the process the country became heavily indebted and had to use much of its export earnings for repayment. Nevertheless, by the mid-1990s, three-fourths of Hungary's trade was with market economies. Germany became Hungary's most important trading partner, followed by Austria and Italy. In 1994 the republic became an associate member of the European Union, and in 1996 it joined the Organisation for Economic Co-Operation and Development.

By the mid-1990s raw materials, semifinished goods, and spare parts represented more than one-third of exports. About one-fourth of exports were industrial consumer goods and nearly another quarter consisted of agricultural and processed food products. Energy, raw materials, and semifinished products constituted more than half of the total imports. One-fifth of imports were industrial consumer goods, and another fifth consisted of machinery and capital goods (such as motor vehicles, tools, and paper).

Transportation. Railways have long been the centre of Hungary's transportation system. By World War I the country had a modern network that was among the densest in Europe, and it continued to expand regularly until the late 1970s, with electrification beginning the previous decade. When industrial production declined during the transition to a market economy, rail transport of goods dropped sharply, accompanied by significant cutbacks in government subsidies that contributed to the deterioration of the railway infrastructure.

By the mid-1990s road haulage made up an increasing percentage of overall transport of goods. Hungary has more than 18,630 miles (30,000 kilometres) of roads, and expansion of the country's highway system is a priority.

Passenger transportation

Buses are the main form of passenger transportation, with long-distance bus passengers outnumbering rail passengers three to one. The number of privately owned automobiles grew rapidly after the early 1980s.

The Danube, the country's only important transportation waterway, is used primarily for international shipping, via the free port of Csepel. Air transport is limited to interna-

tional flights through Ferihegy International Airport near Budapest. MALÉV, the Hungarian national airline, was founded in 1946.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The modern political system in Hungary was essentially autocratic throughout the 19th and 20th centuries. After World War II the Soviet-style system was introduced, with a leading role for the Communist Party, to which the legislative and executive branches of the government and the legal system were subordinated. Other political parties were abolished. In 1948 the Communist Party and the Social Democratic Party merged to form the Hungarian Workers Party, which was reorganized as the Hungarian Socialist Workers Party (MSzMP) after 1956.

Constitutional framework. In 1989 dramatic political reforms accompanied the economic transformation taking place. After giving up its institutionalized leading role, the MSzMP abolished itself and formed the Hungarian Socialist Party (MSzP). In October 1989 a radical revision of the 1949 constitution, which included some 100 changes, introduced a multiparty parliamentary system of representative democracy, with free elections. The legislative and executive branches of the government were separated, and an independent judicial system was created. The revision established a Constitutional Court, elected by the parliament, which reviews the constitutionality of legislation and may annul laws.

The National Assembly

Supreme legislative power is granted to the 386-member unicameral National Assembly, which elects the president of the republic, the Council of Ministers, the president of the Supreme Court, and the chief prosecutor. The main organ of state administration is the Council of Ministers, which is headed by the prime minister. The president, who may serve two five-year terms, is commander in chief of the armed forces but otherwise has limited authority. The right of the people to propose referendums is guaranteed.

Parliamentary elections based on universal suffrage for citizens aged 18 years and over are held every four years. Under the mixed system of direct and proportional representation, candidates may be elected as part of national and regional party lists or in an individual constituency. In the latter case candidates must gain an absolute majority in the first round of the elections, or runoff elections must be held. Candidates on territorial lists cannot be elected if their party fails to receive at least 4 percent of the national aggregate of votes for the territorial lists.

Political parties. About 200 political parties were established following the revision of the constitution in 1989, but only six of them became permanent participants in the country's new political life after the first free elections (1990): the Hungarian Democratic Forum, Alliance of Free Democrats, Independent Smallholders' Party, Christian Democratic People's Party, Federation of Young Democrats, and Socialist Party.

Local government. Hungary is divided administratively into 19 counties (*megyék*), cities, towns, and villages. Budapest has a special status as the capital city (*főváros*) and is further divided into districts. Local representative governments are responsible for protection of the environment, local public transport and utilities, public security, and various economic, social, and cultural activities. Public administration offices, whose heads are appointed by the minister of the interior, supervise the legality of the operations of local governments.

Judiciary. Justice is administered by the Supreme Court, which provides conceptual guidance for the judicial activity of the Court of the Capital City and the county courts and for the local courts. A chief prosecutor is responsible for protecting the rights of citizens and prosecuting acts violating constitutional order and endangering security.

Armed forces. The Hungarian armed forces consist of ground forces, air and air defense forces, a small navy, the border guard, and police. Military service is compulsory for males over the age of 18 (its length varies according to the branch of service but is typically less than one year). The armed forces are not permitted to cross the state frontiers without the prior consent of parliament.

Education. Schools were nationalized in Hungary in 1948. Attendance of the country's extensive system of preschools and kindergartens is not mandatory, but most children attend. Education is free at all levels. It is compulsory from age 6 to 16 and includes an eight-year general education program. Secondary education, which includes both vocational training and preparation for higher education, became virtually universal by the early 1980s. Hungary's numerous institutions of higher educationwere reorganized after World War II. In addition to major institutions such as Loránd Eötvös University in Budapest, Lajos Kossuth University in Debrecen, Janus Pannonius University of Pécs, and Attila József University in Szeged, there are hundreds of specialized schools and colleges. The Central European University, an institution of postgraduate study, primarily serves students from the former Soviet bloc. Severe budget cuts during the transition period resulted in the reintroduction of tuition for higher education. Those cuts also contributed to the reestablishment of private and church-run schools and threatened the existence of the preschool system.

Universities

Health and welfare. Health care improved dramatically under state socialism, with significant increases in the number of physicians and hospital beds in Hungary. By the 1970s free health insurance was guaranteed. Private health care, permitted but limited before the transition period, grew in importance from the early 1990s.

A broad range of social services was provided by the communist government, including child support, extensive maternity leave, and an old-age pension system. This costly welfare system was a heavy burden on the country's budget. At the end of the communist era, Hungary ranked 20th among European nations in terms of per capita GDP, but it was 12th in social insurance spending. Social insurance expenditure, which constituted 4 percent of the republic's GDP in 1950, had risen to one-fifth of the GDP by 1990. The Hungarian system had become one of the most expensive in the world, yet there was considerable resistance to efforts to scale it back. Nevertheless, reform of the system in 1992 made health insurance mandatory but required employers and employees to contribute to that as well as to pension plans.

Housing. Housing shortages were constant in Hungary for years after World War II, despite the million housing units built by the state in urban centres from 1956 to 1985. By the mid-1990s, however, the average number of persons per room in Hungary dropped to roughly one (down from nearly three per room in the immediate postwar period.) Moreover, by the late 1980s electricity was available for nearly the entire population, and running water was available for more than three-fourths of homes. The construction of private homes, which had increased in the 1960s and '70s, constituted more than four-fifths of all construction by the mid-1990s, as housing became part of the market economy.

CULTURAL LIFE

The cultural milieu of Hungary is a result of the diverse mix of genuine Hungarian peasant culture and the cosmopolitan culture of an influential German and Jewish urban population. Both the coffeehouse (as meeting place for intellectuals) and Gypsy music also have had an impact. Cultural life traditionally has been highly political since national culture became the sine qua non of belated nation building from the early 19th century. Theatre, opera, and literature in particular played crucial roles in developing national consciousness. Poets and writers, especially in crisis situations, became national heroes and prophets. Governments also attempted to influence cultural life through subsidy and regulation. During the state socialist era culture was strictly controlled; party interference was influenced by ideological principles, and mass culture was promoted.

Daily life. Genuine traditional culture survived for a long period in an untouched countryside characterized by rootedness. Peasant dress, food, and entertainment, including folk songs and folk dances—the rituals of weddings and Easter and Christmas holidays—continued until the mid-20th century. The drastic (and in the countryside brutal) modernization of the second half of the 20th cen-

tury nearly destroyed these customs. They were preserved, however, as folk art and tourist entertainment.

Everyday life changed dramatically, as did the family structure. Families became smaller, and ties with extended families diminished. The culture also became less traditional. Clothing styles began to follow the international pattern, and traditional peasant dress was replaced by blue jeans. Folk songs are still occasionally heard, but in daily life they have been replaced by modern rock and pop music. Urban culture, especially in the capital city, is highly cosmopolitan and encompasses the tradition of coffeehouse culture. Watching television is a popular pastime.

Cuisine

Hungary's most traditional cultural element is its cuisine. Hungarian food is very rich, and red meat is frequently used as an ingredient. Goulash (*gulyás*), bean soup with smoked meat, and beef stew are national dishes. The most distinctive element of Hungarian cuisine is paprika, a spice made from the pods of chili peppers (*Capsicum annuum*). Paprika is not native to Hungary—having been imported either from Spain, India by way of the Turks, or the Americas—but it is a fixture on most dining tables in Hungary and an important export. Among Hungary's spicy dishes are *halászle*, a fish soup, and *lecsó*, made with hot paprika, tomato, and sausage. Homemade spirits, including various fruit brandies (*pálinka*), are popular. Before World War II, Hungary was a wine-drinking country, but beer has become increasingly prevalent.

Arts. Traditional folk arts either have disappeared or have become mostly commercialized, and political attempts in the 1930s, '50s, and '70s to preserve them basically failed. Beginning in the 19th century, a national high culture emerged, with literature taking a central role. Ferenc Kazinczy, an advocate of Enlightenment ideas, founded a movement of language reform and promoted literature through his high standard of literary criticism. In his view, literature was a nation-sustaining or even nation-creating force. A newly born literary language was cultivated by Mihály Csokonai Vitéz's rococo poetry and plays. Hungarian drama was born with József Katona's tragedy *Bánk Bán*. The brothers Károly and Sándor Kisfaludy established the Hungarian novel, and the first Hungarian language newspaper, *Hazai Tudósítások*, appeared in 1806. Among other important 19th- and early 20th-century literary figures were the poets Sándor Petőfi, János Arany, and Endre Ady. In the *Discovery of Hungary* series of books written in the 1930s, Gyula Illyés, Ferenc Erdei, and others explored the reality of peasant life. In 1956 writers who had been forced to follow the strict rules of Socialist Realism joined the revolt against the regime; the literary and political contributions of Tibor Déry and István Örkény were particulary significant. Notable among the younger generation of writers that followed are Péter Nádás and Péter Esterházy.

Most of the important achievements in Hungarian visual arts and music emerged about the turn of the 20th century. The avant-garde painters Lajos Kassák and László Moholy-Nagy elevated Hungarian painting from second-rate romanticism to international significance. Hungarian music achieved worldwide renown with the composer Béla Bartók, a central figure of early 20th-century culture who influenced future generations of Hungarian composers. Musical life and education in Hungary—including the special schools where the method established by Zoltán Kodály is taught—are on par with those of any nation. Many Hungarian musicians have gained international renown as performers, including the conductors Sir Georg Solti and Antal Dorati, the composer and pianist Franz (Ferenk) Liszt, and the pianists Annie Fischer, Zoltán Kocsis, and András Schiff.

From the 1960s Hungarian motion pictures attracted significant international interest. In particular the parabolic films of Miklós Jancsó and István Szabó helped establish the reputation of Hungarian cinema.

Cultural institutions. Following World War II, high culture that previously had been confined to the upper classes was promoted among the masses. A highly subsidized publishing industry fostered reading: the number of books published increased 10-fold between 1938 and 1988. Reading became a regular habit of a third of the population, and a huge network of more than 15,000 public libraries was established. The main national collections are the Széchényi Library, Ervin Szabó Library, and the libraries of the parliament and the Hungarian Academy of Sciences. Among the most notable of the more than 15,000 museums and roughly 3,000 cultural centres are the National Gallery, the National Museum, and the Museum of Fine Arts (all in Budapest) and the Christian Museum in Esztergom. Museums in Budapest and Pécs house the works of the Hungarian-born artist Victor Vasarely. Government subsidizing of culture virtually ended with the introduction of a market system in the 1990s.

Scholarship

Scholarship and research are both emphasized in Hungary's institutions of higher learning. Following the Soviet model, teaching and research are mutually exclusive. The former is practiced in schools and universities, while the latter is concentrated at the research institute network of the Hungarian Academy of Sciences (established in 1825 and reorganized in 1949). The Academy, with its several dozen research institutes in science, social sciences, and humanities, employs roughly 10,000 researchers and staff.

Hungary has an international reputation for scholarship, with the world's highest per capita rate of Nobel laureates. Because of a lack of funding, however, most of these prizewinners work abroad. Outstanding Hungarian-born scholars include the scientists John von Neumann, Leo Szilárd, Edward Teller, Eugene Wigner, and Albert Szent-Györgyi, the social scientist Karl Mannheim, the economist Karl Polanyi, and the philosopher and literary critic György Lukács. In addition, Pál Turán and Péter Erdös are among the world's most renowned mathematicians. Hungarian scholars also have excelled in the disciplines of linguistics (especially historical linguistics), historiography, and literary history.

Recreation. In addition to their observance of the main religious holidays—Christmas, celebrated as a traditional family festivity, and Easter, characterized by merry village rituals—Hungarians commemorate the labour movement with a major celebration on May 1. Among the traditions of the labour movement celebration is folk dancing, which, like choral singing, also is practiced by peasant ensembles. The Pilis, Mátra, and Bükk mountains are popular vacation destinations for many Hungarians, and a major attraction for the people of Budapest is Margit (Margaret) Island, an urban oasis of gardens and swimming pools in the middle of the Danube.

Hungary has a tradition of success in international sporting competition. It has won a number of world championships and Olympic medals largely because of the overpoliticization of sports in Soviet-bloc nations during the Soviet period. Soccer (association football) is especially popular, and Hungarian athletes also have enjoyed



Lake Balaton from the northwestern shore, near Badacsony, Hung.

Gerd Wagner/B&U International Picture Service, Amsterdam

success in fencing, swimming, table tennis, track and field (athletics), rowing, and weight lifting. Tennis has gained in popularity, especially among the middle class.

Press and broadcasting. Under state socialism the press—about 30 daily newspapers and 1,500 periodicals—was strictly controlled, yet it remained the least restricted in the Soviet bloc. In 1988 press censorship was relaxed. In the first half of the 1990s the number of newspapers increased, but overall circulation declined.

After World War II radio ownership and listening became common. Television appeared only in the late 1950s but soon conquered the country. From the early 1980s almost every household has had a television. There are two main state-run TV channels, and cable and satellite TV are also available. Weekly programming doubled during the 1990s, from about 100 to more than 200 hours. (G.Ba./I.T.B.)

For statistical data on the land and people of Hungary, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

Hungary came into existence when the Magyars, a Finno-Ugric people, occupied the middle basin of the Danube River in the late 9th century AD. Parts of its territory had formed the ancient Roman provinces of Pannonia and Dacia. When Rome lost control of Pannonia at the end of the 4th century, it was occupied first by Germanic tribes, then by Slavs. The subsequent history of Dacia is unrecorded. The central plains had formed the bases of nomadic immigrant peoples from the steppes north of the Black Sea—Huns, Bulgars, Avars—some of whom extended their domination farther afield. The Avars, who dominated the basin in the 7th and 8th centuries, were crushed in about 800 by Charlemagne, whose successors organized the western half of the area in a chain of Slavic vassal "dukedoms." One of these, Croatia, made itself fully independent in 869, while another, Moravia, had openly defied its Carolingian overlord for as long. The Byzantine Empire and Bulgaria exercised loose authority over the south and east of the area.

THE KINGDOM TO 1526

The history of the Magyars

The Árpáds. In 892 the Carolingian emperor Arnulf, attempting to assert his authority over the Moravian duke Svatopluk, called in the help of the Magyars, whose early homes had been on the upper waters of the Volga and Kama rivers; unrecorded causes had driven them southward into the steppes, where they had adopted the life of peripatetic herders. In the 9th century they were based on the lower Don, ranging over the steppes to the west of that river. They then comprised a federation of hordes, or tribes, each under a hereditary chieftain and each composed of a varying number of clans, the members of which shared a real or imagined blood kinship. All clan members were free, but the community included slaves taken in battle or in raids. There were seven Magyar hordes, but other elements were part of the federation, including three hordes of Turkic Khazars (the Kavars). Either because of this fact or perhaps because of a memory of earlier conditions, this federation was known to its neighbours as the On-Ogur (literally, "Ten Arrows"), from the Slavic pronunciation of which the name "Hungarian" is derived.

In 889 attacks by a newly arrived people called the Pechenegs had driven the Magyars and their confederates to the western extremities of the steppes, where they were living when Arnulf's invitation arrived. The band sent to Arnulf reported back that the plains across the Carpathian Mountains would form a suitable new homeland that could be easily conquered and defended from the rear. Having elected as their chief Árpád, the leader of their most powerful tribe, the Magyars crossed the Carpathians en masse, probably in 896, and easily subjugated the peoples of the sparsely inhabited central plain, their first place of settlement. They destroyed the Moravian empire in 906 and in the next year occupied Pannonia, having defeated a German force sent against them. They were then firmly established in the whole centre of the basin, over which their tribes and their associates distributed themselves, Árpád taking the central area west of the Danube for his own tribe. The periphery was guarded by outposts, which were gradually pushed forward, chiefly to the north.

The Christian kingdom. During the next half century the Magyars were chiefly known in Europe for the forays they made across the continent, either as mercenaries in the service of warring princes or in search of booty for themselves—treasure or slaves. Terrifying to others, their mode of life was not always profitable. Their raiding forces suffered a number of severe reverses, culminating in a disastrous defeat at the hands of the German king Otto I in 955 at the Battle of Lechfeld, outside Augsburg (now in Germany). By that time the wild blood of the first invaders was thinning out and new influences, in particular Christianity, had begun to operate. Both the Eastern and Western churches strove to draw them into their orbits. They had established pacific, almost friendly relations with Bavaria. The decisive step was taken by Árpád's great-grandson Géza, who succeeded to the hereditary chief leadership in 972 and reestablished its authority over the tribal chiefs. In 973 he sent an embassy to the Holy Roman emperor Otto II at Quedlinburg (now in Germany), and in 975 he and his family were received into the Western church. In 996 his son, Stephen (István), married Gisella, a Bavarian princess.

Magyar defeat at Lechfeld

Stephen I (997–1038) carried on his father's work. With the help of Bavarian knights, he crushed his rivals for the headship. Applying to Pope Sylvester II, he received the insignia of royalty from the papacy and was crowned king on Christmas Day, 1000. The event was of immeasurable importance, for not only did Hungary enter the spiritual community of the Western world but it did so without having to recognize the political suzerainty of the Holy Roman Empire. This was possible because Sylvester, who extended papal protection to Hungary, held great sway with the emperor, Otto III, who had once been his pupil. Stephen then effected the conversion of his people to Christianity, establishing a network of archiepiscopal and episcopal sees. Later he crushed the surviving disputants of his authority—notably the Kavars—and, furthering his father's work, organized his state on a system that was to remain for many centuries the essentially unaltered basis of Hungary's political and social structure. The tribes, as units, disappeared, but the fundamental social stratification was not altered. The descendants in the male line of the old conquerors and elements later equated with them remained a privileged class, answerable only to the king or his representative and entitled to appear in general assemblage. Their lands—which at this time, since the economy was mainly pastoral, were held by clans or subclans in semicommunal ownership—were inalienable, except for proved delinquency, and free of any obligation. The only duty required by the state of members of this class was that of military service on call. They were allowed to retain their slaves, although Stephen freed his own. All land not held by this class—then more than half the whole—belonged to the crown, which could donate it at will. The nonservile inhabitants of these lands—*e.g.*, descendants of the pre-Magyar population, manumitted slaves, and invited colonists—were subjects of the crown or of the local landholder.

Conversion to Christianity

The whole of this land was divided into counties (*megyék*), each under a royal official called an *ispán* (later *fő* [head] *ispán*), who represented the king's authority in it, administered its unfree population, and collected the taxes that formed the national revenue, central and local. Each *ispán* maintained at his *vár*(fortress) an armed force composed of freemen who took service under him or of persons freed by the king. In Stephen's day there were about 45 such counties.

The early kings. Once St. Stephen (he was canonized in 1083) established his rule, his authority was rarely questioned. He fought few foreign wars and made his long reign a period of peaceful consolidation. But his death in 1038 was followed by a period of internecine dynastic jealousies and intrigue (including murder) that lasted more than a century and a half, with few intervals of stability, before the 30-year reign of Andrew II(1205–35).

A period of disputed succession

Consolidation and expansion. These royal disputes

caused Hungary much harm. Claimants to the throne often invoked foreign help, for which they paid in political degradation or loss of territory; both Peter (1044–46) and Salamon, the son of Andrew I, did homage to the Holy Roman emperor for their thrones; and Samuel Aba's war against Peter's protectors cost Hungary its previous territories west of the Leitha River, while the wars of the 12th century cost it areas in the south. The uncertainty delayed political consolidation, and even Christianity did not take root easily: there was a widespread pagan revolt in 1047 and another in 1063.

Yet the political unity of the country and the new faith somehow survived the earlier troubles, and both were firmly established by Ladislas I (1077–95), one of Hungary's greatest kings, and by Coloman (Kálmán; 1095–1116), who, despite the fact that he had his brother and nephew blinded to secure the throne for his own son, was a competent and enlightened ruler.

Meanwhile, many factors worked for Hungary. After Austria had grown big at the expense of the imperial authority, most of Hungary's neighbours were states of approximately the same size and strength as itself, and the Hungarians lived with them on terms of mutual tolerance and even friendship. The steppes were quiet: the Cumans (Hungarian: Kun), after destroying the Pechenegs there, did not try to go farther, and, after two big raids had been successfully repelled by Ladislas I, they left Hungary in peace. This allowed Hungary to extend its effective frontiers to the Carpathian crest in the north and over Transylvania. Magyar advance guards pushed up the valleys of both areas and were reinforced in the Szepes area and in central Transylvania by imported colonies of Germans (usually called Saxons). In the meantime, colonies of Szeklers (Székely), a people akin to the Magyars who had preceded the latter into the central plains, were settled behind its eastern passes. The county system was extended to both areas, although with modifications in Transylvania, where the Saxons and Szeklers constituted free communities and the whole was placed under a vaivode, or governor. In the south Ladislas I occupied the area between the Sava and Drava rivers; Coloman assumed the crown of Croatia, which then included Bosnia and northern Dalmatia, although this remained a separate "Land of the Hungarian Crown," over which a governor (ban) acted as deputy for the king.

Imported colonists

In the interior, too, natural growth and continued immigration swelled the population, which by 1200 had risen to the then large figure of some two million. The rulers of this big, populous state were now important men. After Ladislas' day, German claims to suzerainty over Hungary ceased. In the 12th century the country intervened in its neighbours' affairs as often as they did in Hungary's. Béla III (1173–96), who married a French princess, Margaret Capet, had revenues roughly equal to the income of the king of France. He owned half the land of the kingdom outright and held monopolies of coinage, customs, and mining. While the income of the early kings had been mostly in kind, half of Béla's income was cash, coming from royal monopolies and taxes paid by foreign settlers.

Social and political developments. Meanwhile, the pattern of Hungarian society had been changing. The population of the free class, or "nobles," although frequently reinforced by new admissions to its ranks, probably hardly increased in absolute terms, and certainly grew far less than the unfree population: from perhaps half the total population in 896, they had been reduced to about one-eighth by 1200. Further, as the economy became agricultural, the old clan lands dwindled until only pockets remained. Where the rest had been and in large parts of the old crown lands, which improvident donations had greatly reduced, the land was held in the form of individual estates. The owner of each of these estates was master of the unfree population on it: the nobles had, to a large extent, become a landed oligarchy. Some individual estates were very large, and their owners had come to constitute a "magnate" class, not yet institutionalized or legally differentiated from their poorer co-nobles but far above them in wealth and influence. Although slavery had practically disappeared, the non-nobles were still a "subject" class. Many of them, including the burghers of the towns (most of which were German foundations) and members of such communities as the Saxons and Szeklers, were protected by special charters and personally free, but even they stood, politically, outside the magic ring of the *natio Hungarica*.

As a result of Béla's marriage to the sister of the French king, the Hungarian court became a centre of French knightly culture. Western dress and translations of French tales of chivalry appeared. A royal notary, known to future generations as "Anonymous," wrote the history of the conquest of Hungary. The first known work in the Hungarian language, the "Halotti beszéd" ("Funeral Oration"), was part of the otherwise Latin-language Pray-codex written in the early 1190s. Béla also followed a Western model in introducing written documentation of government administrative authority. Moreover, monasteries served as public notaries from the end of the 12th century.

In addition to tents and wooden structures, stone buildings (mostly churches) appeared in the permanent settlements. The cathedral of Pécs, Pannonhalma Apátság (a Benedictine abbey), and the royal palace at Esztergom were the first examples of early Gothic architecture.

Throughout these developments the country had remained an absolutist patrimonial kingship. The king maintained a council of *optimates* (aristocrats), but his prerogatives were not restricted and his authority remained absolute. A strong king, such as Béla III, could always curb a recalcitrant magnate by simply confiscating his estate. Only the follies and extravagances of the feckless Andrew II evoked a revolt, culminating in 1222 in the issue of the Golden Bull (the eastern European equivalent of the Magna Carta), to which every Hungarian king thereafter had to swear. Its purpose was twofold: to reaffirm the rights of the smaller nobles of the old and new (*servientes regis*) classes against both the crown and the magnates and to defend those of the whole nation against the crown by restricting the powers of the latter in certain fields and legalizing refusal to obey its unlawful commands (the *jus resistendi*). Andrew had done much harm by dissipating the royal revenues through his extravagances and by issuing huge grants of land to his partisans. The royal estate gradually melted away as the *ispáns* and knights became the hereditary owners of the land. Leading aristocratic families like the Abas and Csáks became the unchallenged rulers of large parts of the country.

Issuance of the Golden Bull

The Mongol invasion: the last Árpád kings. Andrew's successor, Béla IV (1235–70), began his reign with a series of measures designed to reestablish royal authority, but his work was soon interrupted by the Mongol invasion. In the spring of 1241 the Mongols overran the country and, before they left it, a year later, had inflicted ghastly devastation. Only a few fortified places and the impenetrable swamps and forests escaped their ravages. The country lost about half its population, the incidence ranging from 60 percent in the Alfold (100 percent in parts of it) to 20 percent in Transdanubia; only parts of Transylvania and the northwest came off fairly lightly. Returned from Dalmatia, where he had taken refuge, Béla, whom his country not unjustly dubbed its second founder, reorganized the army, built a chain of fortresses, and called in new settlers to repopulate the country. He paid special attention to the towns. But he was forced to give some of the magnates practically a free hand on their own estates, and a few families rose to near-sovereign local status. Further, one group of immigrants, a body of Cumans who had fled into Hungary before the Mongols, proved so powerful and so turbulent that to ensure their loyalty Béla had to marry his son, Stephen V, to a Cuman princess. The king attempted to counterbalance the power of the magnates by creating his own army, partly from the Cumans. A newly created "conditional" nobility comprising enobled soldiers and settlers who gained land for military service strengthened the ranks of the lesser nobility. The system of royal estates and judicial power was thereafter transformed in an assembly in which nobles represented their counties.

Rebuilding under Béla IV

Stephen died two years after his father's death, after which the country passed under the regency of his widow, the "Cumanian woman," whom the Hungarians detested.

Her son was assassinated and left no legitimate heir, and claims to the throne were made through the female line of the Árpáds. A male heir was found in Italy, and, although the young man's claim to the throne was impugned, Andrew III proved a wise, capable king. With his death in 1301, however, the national dynasty became extinct.

A new Western-style feudal socioeconomic system had emerged in Hungary but it had yet to take root. During the last third of the 13th century Hungarian assimilation into Europe was threatened by the ongoing conflicts between various baronial factions. Moreover, Hungary was still the destination of migrating pagan tribes and the focus of barbarian attacks, and it continued to exhibit the features of a country on the borders of Christian feudal Europe.

Hungary under foreign kings. The extinction of the old dynasty entitled the nation to choose its successor; but the principle of the blood tie was still generally regarded as determinant, and all the candidates for the throne—Wenceslas of Bohemia, Otto of Bavaria, and Charles Robert of the Angevin house of Naples—based their claims on descent from an Árpád in the female line. But all three claimants were foreigners. From that time until its extinction, the kingship of Hungary was invariably—with two exceptions—held by a foreigner, nearly always by one occupying simultaneously at least one foreign throne. This could be to the advantage of Hungary if the king used the resources of those thrones in its service, but he could, alternatively, neglect and exploit Hungary in his other interests and use his power to crush its national freedoms and institutions. Securing the advantages of foreign rule while escaping its dangers was the abiding dilemma—seldom successfully resolved—of Hungarian history.

The Angevin kings. Charles Robert of Anjou (1308–42), still a child when his supporters won out, had no foreign throne and grew up a true Hungarian. He was also a capable man. After reaching manhood, he crushed the most rebellious of the "kinglets" and won over the rest; after this his rule was unquestioned and peaceful at home. The international situation, with Germany distraught by the power struggle between empire and papacy, the Mongolian Tatars grown passive, and the power of Byzantium in full decay, was again favourable to the states of the "middle zone" of eastern Europe and the Balkans; it is no accident that Poland, Hungary, Bohemia, and Serbia all look on the 14th century as their golden age. As this situation favoured Hungary's neighbours, as well as itself, Charles Robert's attempts at expansion were only moderately successful. In the Balkans he made Bosnia his friend and client but lost Dalmatia to Venice and other territories to Serbia and the newly emerged *vaivody* (province) of Walachia. But he drove Czech and Austrian marauders out of the land and on the whole preserved friendly relations with Austria, Bohemia, and Poland.

Charles's son, Louis (Lajos) I (1342–82), the only Hungarian king on whom his country has bestowed the name "Great," built on his father's foundations. Keeping peace with the West, he repaired his father's losses in the south and surrounded his kingdom with a ring of dependencies



Hungary in 1360.

over which Hungary presided as *archiregnum* (chief kingdom) in the Balkans, on the lower Danube, and in Galicia. In 1370 he also ascended the throne of Poland, by virtue of an earlier family compact.

Angevin prosperity

Both Angevin kings owed much to the wealth they derived from the gold mines of Transylvania and northern Hungary, some 35 to 40 percent of which went to the king, enabling him to maintain a splendid court. Spared for two generations from serious invasion or civil war, the country blossomed materially as never before. The population rose to three million and the country contained 49 royal free boroughs, more than 500 smaller towns, and some 26,000 villages. The economy was still mainly rural, but the crafts prospered, trade expanded, and the arts flourished.

The life of the court and the daily life of cities borrowed from western European societies. German settlers and burghers in the cities and the clergy became the main agents of Western culture. The Dominicans built 25 monasteries by the early 14th century and established a theological school in Buda. The Franciscans also established monasteries, as did the Cistercians, Premonstratensians, and Paulines. Romanesque style dominated architecture until the ascendancy of Gothic design in the late 13th century. Cities built impressive churches, such as the Church of Our Blessed Lady (now better known as the Matthias Church) in Buda. The palace of Visegrád, the royal castles of Zólyom and Diósgyör, the miniatures of the Illuminated Chronicle (1360) and the St. George statue in Kolozsvár (1373), together with the earliest codex predominantly in Hungarian (1370) and the finest example of Hungarian poetry, *Omagyar Mária-siralom* ("Old Hungarian Lament of the Virgin Mary," about 1300), testify to the spread of western European culture. The first universities were established in Pécs (1367) and Obuda (1395), though they were short-lived. Yet, in spite of its advancement Hungary remained a less-developed borderland of Europe.

The rule of the two Angevin kings was essentially despotic, although enlightened. They introduced elements of feudalism into the political, and especially the military, system: each lord was responsible for maintaining his own armed contingent (*banderium*). But the magnates were held firmly in check, and Louis reaffirmed the rights and privileges of the common nobles and carried further a process, begun in the previous century, under which the counties were developing from "royal" into "noble" institutions, each still under a royal official but administered with a wide measure of autonomy by elected representatives of the local nobility. He also standardized the obligations of the peasants at the figure of one-ninth of their produce to the lord, a tithe to the church, and a house tax (*porta*) that went directly to the state.

Sigismund of Luxembourg. The benefits of Louis's rule would have been far greater still had he not wasted much money and many lives on endeavours to secure the throne of Naples for his nephew. His foreign acquisitions served his personal glory more than they did the real interests of his country; much of the imposing edifice collapsed when he died, leaving as heirs only two daughters. Louis had designated the elder, Maria, to succeed him on both his thrones, but the Poles refused to continue the union. They accepted the younger daughter, Hedvig (Polish: Jadwiga), as queen but married her to Jogaila (Polish: Jagiełło) of Lithuania. The Hungarians crowned Maria, whose husband, Sigismund of Luxembourg, became her consort in 1387 and after her death eight years later ruled alone until his own death in 1437. Under Sigismund, matters took a sharp turn for the worse, although he did much for the arts and commerce and, above all, for the towns. Also, like Andrew II, he promoted Hungarian political institutions by creating the need for them. The principle that the consent of representatives of the privileged classes, assembled in the Diet, was necessary for the grant of any subsidy or additional taxation, and even, later, for any legislation, dates from his reign, being made necessary by his extravagance and arbitrariness. His frequent and prolonged absences gave rise to the peculiar Hungarian institution of the palatine, an officer elected jointly by the king and the nation, who represented the king during his absences and also acted as intermediary between him

and the nation. But these were only palliatives against bitterly felt abuses. The nation, besides hating him for the cruelty he showed at the outset of his reign against the supporters of a rival, resented the absenteeism of his later years, when he occupied himself chiefly with imperial and Bohemian affairs (he was elected German king in 1411 and became titular king of Bohemia in 1419). There was much discontent among the peasants, who were subjected to heavy exactions by the crown and by their masters, the unrest being aggravated by the spread of radical Hussite religious doctrines from Bohemia; there were serious revolts in northern Hungary and Transylvania. Above all, there was the growing danger from the Ottoman Turks, who, though they had already taken Bosnia from Louis, could not threaten Hungary proper while Serbia still stood. But in 1389 the power of Serbia was broken at the Battle of Kosovo, and the danger became urgent. Sigismund organized a crusade that was disastrously defeated at Nicopolis in 1396. Timur (Tamerlane) gave Europe a respite by his attack on the Turkish rear, but the advance was resumed in 1415. Walachia submitted in 1417; thereafter Transylvania and southern Hungary suffered repeated raids.

The growing Ottoman threat

János Hunyadi and Matthias Corvinus. The Ottoman sultan Murad II was preparing a grand assault on Hungary when Sigismund died in 1437, leaving as his heir a daughter, who was married to Albert V of Austria. The country accepted Albert as Sigismund's successor, but only on condition that he not become Holy Roman emperor or reside abroad without permission of the estates. Albert set about organizing the country's defenses but died in 1439, leaving his widow with an unborn child. To avoid an interregnum and a minority, perhaps with a queen, the country elected Władysław III of Poland as king; within two years of his death in battle against the Turks in 1444, they accepted Albert's son, Ladislas V (called Ladislas Posthumus), appointing as his guardian and governor of Hungary the great general János Hunyadi, who had been repelling the renewed Turkish attacks. Hunyadi kept up the defense under increasing difficulties, constantly thwarted by jealous magnates and harassed by the Czech condottiere Jan Jiskra, while Frederick III (first of the Habsburg emperors) encroached on the western provinces.

Hunyadi died in 1456 after the recapture of Belgrade from the Turks. Ladislas' maternal uncle, Ulrich of Cilli, aware of the country's devotion to Hunyadi, had the latter's elder son assassinated and his younger son Matthias (Mátyás) Corvinus imprisoned in Prague. Ladislas V himself died suddenly a year later. The country was tired of foreign rule and its agents, and on Jan. 24, 1458, a great concourse of nobles acclaimed Matthias king. He was brought to Buda and crowned amid nationwide rejoicing.

The only national king to reign over all Hungary after the Árpáds, Matthias Corvinus has been seen through something of a golden haze by historians. A true Renaissance prince, he was a fine natural soldier, a first-class administrator, an outstanding linguist, a learned astrologer, and an enlightened patron of the arts and learning. His collections of illuminated manuscripts, pictures, statues, and jewels were famous throughout Europe. Artists and scholars were welcomed at his court, which could vie in magnificence with any other on the continent. Sumptuous buildings sprang up in his capital and other centres.

Politically, too, he represented the ideas of the Renaissance. He listened to his council, convoked the Diet regularly, and enlarged the autonomous powers of the counties. But at heart he was a despot; his real instruments of government were his secretaries, men picked by himself, usually young and often of humble origin. His rule was in the main an efficient and a benevolent one. He simplified and improved the administration and the laws, enforcing justice with an even hand. The debit side of his rule was the increased taxation imposed by him for his administrative innovations, his collections (which cost his subjects vast sums), and, above all, the mercenary standing army, 30,000 strong (largely composed of the defeated Hussites and known after its commander, "Black John" Haugwitz, as the Black Army), which he kept as part of the royal *banderium* for use against enemies, at home and abroad.

Matthias' despotism

At first he had much need for such a force; although the Turks were quiescent for a decade, there were discontented magnates, and the Czechs and the Austrians were unquiet neighbours. But, after Matthias had crushed, expelled, or bought off these enemies, had built a chain of fortresses along the southern frontier, and had even reestablished a nominal and, in practice, worthless suzerainty over Bosnia, Serbia, Walachia, and Moldavia, he let himself be drawn into an ever-widening circle of campaigns against Bohemia and Austria. In 1469 he made himself master of Moravia, Silesia, and Lusatia, with the title (although this was also borne simultaneously by George of Podebrady) of king of Bohemia, and in 1478 he forced Frederick to cede him Lower Austria and Styria. He argued that his neighbours were untrustworthy and that he could not organize the great crusade against the Turks without the resources of the imperial and Bohemian crowns. But his subjects were unconvinced, and in 1470 a party actually conspired to replace him with a Polish prince. This enterprise collapsed, and Matthias entered on a complex transaction with the new emperor, Maximilian I, under which his illegitimate son John (he had no legitimate issue) was to marry Maximilian's daughter in return for re-cession of the Austrian provinces and Maximilian's recognition of John. But on May 6, 1490, while on his way to the meeting that should have sealed the bargain, Matthias died suddenly, and the whole enterprise collapsed.

Both Sigismund and Matthias attempted to balance baronial power by strengthening the cities, but they were only partly successful. In contrast with western Europe, urbanization remained moderate, with the development of walled cities lagging behind that of western European counterparts. The number of guilds was limited, and the structure of foreign trade reflected economic backwardness: nearly four-fifths of imports consisted of textiles and about one-eighth of metalware. Exports consisted almost entirely of cattle and wine. The most important aspect of urbanization was the rapid growth of agricultural towns. Instead of the approximately 50 families that made up the 20 to 30 *portae* (taxable units) of the typical village, these oversize peasant settlements (*mezővárosok*) had as many as 500 *portae*. Moreover, the number of these settlements increased from about 300 in the mid-15th century to about 800 in the early 16th century.

The Jagiellon kings: national decay. The magnates, who did not want another heavy-handed king, procured the accession of Vladislas II, king of Bohemia (Uászló II in Hungarian history), precisely because of his notorious weakness—he was known as King "Dobre" (meaning "Good" or, loosely, "OK") from his habit of accepting with that word every paper laid before him. The emperor Maximilian contented himself with reoccupying his lost provinces and establishing a sort of paternal patronage over Hungary. This was consolidated in 1515 by an agreement under which Vladislas' son, Louis, married Maximilian's granddaughter Mary, while Louis's sister, Anne, married Maximilian's grandson Ferdinand, who was to succeed to Louis's thrones if Louis died without an heir. The agreement was made without the consent of the Hungarians, and the Diet in 1505 passed a resolution never again to accept a foreign king. The candidate of the "national party" was János Zápolya, vaivode of Transylvania.

Meanwhile, the magnates had disbanded the Black Army (without replacing it) and allowed the country's fortresses to fall into disrepair. Vladislas was their helpless prisoner; he could make no decision without their consent, and his revenues were looted so ruthlessly that he was reduced to selling Matthias' collections. Nearly all of Matthias' reforms were canceled, and the peasants were oppressed grievously. In 1514 there was a peasant uprising that, unlike those that had preceded it, spread nationwide. It was sparked by the call for a crusade against the Ottomans by the papal legate of eastern Europe, Archbishop Tamás Bakócz. Some 20,000 men gathered near Buda in the spring and, led by a Szekler soldier, György Dózsa, moved on the southern border. The rebellious, antilandlord sentiment of these "crusaders" became apparent during their march across the Great Alfold, and Bakócz canceled the campaign. The peasant leaders not only refused to obey

Peasant rebellion

this order when it reached them in late May but also confronted and defeated the noble's army and burned castles for two months. By mid-July, however, the peasants had been defeated and Dózsa tortured and murdered. The peasants were condemned to perpetual servitude, and their right to free migration was abolished. Although this law was not immediately enforced, it served as the basis for the preservation of serfdom for centuries to come.

Renewed Ottoman threat

When Vladislas died in 1516, his nine-year-old son was proclaimed king as Louis II. The defenses of the kingdom worsened, and in 1521 the new Ottoman sultan, Süleyman I the Magnificent, demanded tribute from Louis; when the demand was rejected, Süleyman took Belgrade. Suddenly alive to the Turkish danger, the magnates voted to reestablish a standing army, but nothing was done to raise it, since each rival faction tried to put the burden of its upkeep on the others. Appeals for help from abroad met with little response. In 1526 the sultan advanced into Hungary. A general call to arms was proclaimed, but the most important forces—those from Transylvania and Croatia—were late in obeying it. Louis, with a force of 16,000 men, moved down the Danube in August and attacked the Turks at Mohács. The Hungarian army, heavily outnumbered, was almost annihilated. Louis himself drowned during his flight. Unable to believe that the pitiful army that had met him was Hungary's national army, the sultan advanced with extreme caution. He occupied Buda on September 10 but returned across the Danube by the end of October, taking with him more than 100,000 captives.

THE PERIOD OF PARTITION

Since the sultan had not meant to remain in Hungary, the disaster of Mohács might have been overcome had the king not perished; but, as it was, both Zápolya and Ferdinand of Habsburg (later Holy Roman emperor as Ferdinand I) claimed the throne, and each had supporters in the country. After each had failed to drive his rival out, Zápolya appealed to the sultan, who installed him in Buda, thus limiting Ferdinand's rule to the western third of the country. By a secret agreement—mediated in 1538 by Zápolya's adviser, György Martinuzzi ("Friar George")—Ferdinand was to succeed Zápolya when the older man died. The agreement was upset when, just before Zápolya died, his wife bore a son whom the national party recognized as king. The sultan recognized the infant as king but, more importantly, as his own vassal. Extending his occupation of Buda, the sultan incorporated a great wedge of central and southern Hungary in his own dominions. Ferdinand had to conclude a truce and to pay a tribute in return for recognition of his de facto rule over the territory then held by him. The transaction was completed when in 1566 the sultan formally declared Transylvania an autonomous principality under his own suzerainty; two years later Ferdinand's successor, Maximilian II, was forced to recognize this arrangement and to accept the reduction of Royal Hungary (*i.e.,* that part of Hungary that came under Habsburg rule) to the western fringe of the country, the northwestern mountains, and Croatia. From that time, ruling princes of Transylvania such as István Báthori and later Gábor Bethlen introduced mercantilist economic policies and generated prosperity.



The partition of Hungary in 1568.

The "age of trisection" was the bleakest in all Hungarian history. Fighting and slave raiding, which went on even in times of nominal peace, reduced the whole south of the country to a wasteland occupied by only a few seminomadic Vlach herdsmen; villages disappeared and fields reverted to swamp and forest. Behind the new frontier the population was partially preserved to supply the garrisons, but the old landholders were replaced by Turkish officials and soldiers who, since their fiefs were not hereditable nor even always long-term, exploited the wretched cultivators. Conditions were relatively tolerable only in those *kazalar* (districts) managed directly by the Ottoman government. In those districts, most of which lay along the banks of the Tisza River, the people flocked into the great "village towns" that are still a feature of the area. There they enjoyed a measure of protection; but the country between these towns was left virtually empty.

The Turks left Transylvania relatively unmolested. Martinuzzi devised a constitution based on earlier institutions, consisting, under the prince, of representatives of the three "nations": the Hungarian nobles of the counties, the Saxons, and the Szeklers. Transylvania was also spared internecine religious strife, when the Roman Catholic, Calvinist, Lutheran, and Unitarian churches agreed to coexist on a basis of equal freedom and mutual toleration. The Greek Orthodox faith of the Romanians, who constituted the rest of the population, was only "tolerated," since the Romanians as such, or even their nobles, did not constitute a "nation."

Religious toleration

Royal Hungary and the rise of Transylvania. In the first years after his accession, Ferdinand still hoped to bring the whole kingdom under his rule. He respected its constitution and its institutions and convoked the Diet regularly. But his hopes faded, and, after his succession to the imperial crown in 1558, Royal Hungary became no more than a small outlying annex of his mighty dominions. As it was also an exposed one, without the resources to defend itself, Ferdinand and his successor, Maximilian II, organized a chain of fortresses, mostly garrisoned by German troops, and a defensive "military frontier," inhabited by Serb and Vlach refugees from the Balkans and administered from Vienna. The Hungarians complained that they were being ruled and exploited as a subject people by foreigners, while Vienna looked on them as truculent rebels. Matters grew worse when Maximilian was succeeded by the mentally unbalanced Rudolf II, whose advisers hated Hungary and its traditions; and a religious conflict supervened on the constitutional dispute, for in the preceding half century the Reformation had swept over Hungary.

Religious antagonism played an important part when war between the empire and the Turks broke out again in 1591. In the so-called Fifteen Years' War, imperial troops entered Transylvania, and their commander, George Basta, behaved there (and in northern Hungary) with such insane cruelty toward the Hungarian Protestants that a Transylvanian general, István Bocskay, formerly a Habsburg supporter, revolted. His army of wild herdsmen (*hajduks*) drove out Basta, and in June 1606 Bocskay concluded with Rudolf the Peace of Vienna, which left him prince of an enlarged Transylvania and also guaranteed the rights of the Protestants of Royal Hungary. He then mediated the Peace of Zsitvatorok (November 1606) between the emperor and the sultan, which kept the territorial status quo but relieved the emperor of his tribute to the sultan.

These two treaties ushered in a new era. The balance of power began to shift from the Turks toward Transylvania, which entered a half century of prosperity. A scramble for power followed Bocskay's death (1606), but in 1613 the Sublime Porte (the Ottoman government) imposed the election of Gábor Bethlen (1613–29), who proved the most famous of all the princes of Transylvania. At home Bethlen's rule was thoroughly despotic: through his monopoly of foreign trade and his development of the principality's internal resources, he almost doubled his revenues, devoting the proceeds partly to the upkeep of a sumptuous court, partly to the maintenance of a standing army. Keeping peace with the Porte, he often intervened against the emperor in the Thirty Years' War (1618–48) and safeguarded the rights of the Protestants in Royal

Reign of Gábor Bethlen

Hungary. Under the Treaty of Nikolsburg (Dec. 31, 1621) he retained the title of prince of Transylvania and Hungary (the Porte vetoed his acceptance of the Hungarian crown) and gained a big extension of the principality and a duchy in Silesia, besides further guarantees for the Protestants of Royal Hungary. When Bethlen died suddenly in 1629, his subjects abolished most of his internal reforms, but his successor, György Rákóczi I, maintained the international position of Transylvania, which figured as a sovereign state in the Peace of Westphalia (1648), ending the Thirty Years' War. This Transylvanian support for the Protestants in Royal Hungary, as well as the divisions prevailing among their own members, prevented the Habsburgs from enforcing the Counter-Reformation in Hungary as early and as fully as they did in Austria and Bohemia. Nevertheless, the genius of the cardinal-primate Péter Pázmány won over for Roman Catholicism the majority of the local magnates, who came to form a party attached to the Habsburg cause, which was the more influential because they now formed a separate "table" of the Diet. The nation was thus divided not only between Transylvania and the west but also between the Roman Catholic magnates and their subjects on the one hand and the smaller landowners, many of whom remained Protestants, on the other. In religious matters, the Hungarian Catholic nobles were no more tolerant toward their Protestant fellow countrymen than were the emperor's own German and Czech advisers, although they were not willing to acquiesce in the political centralization championed by the latter.

War and liberation. The Turkish occupation of central Hungary remained a volatile issue, for every Hungarian resented the Habsburgs' policy of leaving the Turks unmolested while pursuing ambitious objectives in the west. This powder keg erupted in 1657 when György Rákóczi II of Transylvania, who had succeeded his father in 1648, allowed the prospect of obtaining the crown of Poland to seduce him into sending across the Carpathians an expeditionary force, which was annihilated by Tatars. The grand vizier Köprülü Mehmed Paşa, the architect of the Porte's renaissance, led a force against Transylvania and installed a new puppet prince. Emperor Leopold sent a force against the Turks; but, although the Austrian general Raimondo Montecuccoli defeated the Turks at Szentgotthárd on Aug. 1, 1664, the subsequent Peace of Vasvár recognized all the sultan's gains.

Reassertion of Turkish power

Now even the highest magnates of Royal Hungary plotted to expel the Habsburgs with Turkish and French help, but the conspiracy was betrayed, and Vienna took its revenge. Nobles were executed or lost their estates, and Protestant pastors were sentenced to be galley slaves. In 1673 the constitution was suspended and Hungary placed under a directorate. A young Transylvanian, Imre Thököly, led a revolt that forced Leopold in 1681 to restore the constitution and revoke many of his harshest measures. The Porte, encouraged by Thököly's successes, sent into Hungary a vast army that in 1683 reached Vienna. But the tide ebbed as swiftly as it had advanced. Vienna was relieved and the Turks routed, and the imperial general Prince Eugene of Savoy led a series of campaigns in which all western and central Hungary, with Buda, were cleared of the Turks in 1686. Transylvania was liberated in the years following. By the Treaty of Carlowitz (January 1699), the sultan relinquished all of Hungary except the corner between the Maros and Tisza rivers. (This area was ceded in 1718 but kept until 1779 under Austrian administration as the Bánát of Temesvár.) The military frontier (progressively extended) was kept under a similar regime, and Transylvania was organized as a separate "principality."

HABSBURG RULE, 1699–1918

Habsburg rule to 1867. The emperor, not Hungary, was the victor, for the retreating Turks and the advancing armies of the so-called liberators ravaged the country. In 1687 Leopold reconfirmed the constitution subject to Hungary's acceptance of his dynasty in the male line and to the abolition of the *jus resistendi* conceded under the Golden Bull of 1222, but the government that followed was another cruel Vienna-centred dictatorship. In 1703 this provoked another rebellion, led by Ferenc Rákóczi II (Thököly's stepson); after eight years of indecisive and fruitless fighting, peace was established by the Treaty of Szatmár (April 1711). On paper, this did little more than confirm what had been agreed in 1687, but the new king, Charles III (Emperor Charles VI), genuinely wanted peace with Hungary, and the worst abuses were now ended.

Charles III and Maria Theresa. Charles's chief concern was to secure the acceptance in Hungary of the Pragmatic Sanction, the imperial decree by which his daughter Maria Theresa was to inherit his dominions. After the Diet accepted the Pragmatic Sanction in 1723, Charles convoked the body only once more, and Maria Theresa, after her coronation in 1740, only twice more—each time to ask for money. Her rule, like her father's, was essentially autocratic. She was severe toward the Protestants, and she allowed her advisers to exclude Hungary from the subsidized industrialization that was bringing wealth to other parts of her dominions. Internal tariff barriers were introduced between the hereditary provinces and Hungary. Imports to Hungary from outside the empire were hindered by high tariffs, but customs for "imports" from Austria and Bohemia were very low. Hungary's exports were all but banned to non-Habsburg lands, and only those agricultural and raw materials that were required in the western part of the monarchy received preferential treatment. Hungary became more dependent on, and subordinate to, Austria than before. Agriculture received some incentives, but the road to industrialization was blocked. Lacking modern credit, entrepreneurial attitude, and strong urban markets, Hungary, unlike Austria and Bohemia, was prevented from entering the preindustrial age.

The Pragmatic Sanction

Maria's rule was not more than severe, even toward the Protestants. Toward the magnates, on whom she lavished many favours, it was positively benign, and she respected the most cherished liberty of the lesser nobles: their exemption from taxation. Exhausted by so many wars and rebellions, the country asked for nothing more, contenting itself with the supreme blessing that her rule brought it an uninterrupted peace that enabled the population to grow once again and the material ravages to be repaired. But a lethargy descended on the country. Political life sank to the parish-pump level, and the towns stagnated. The peasants, into whose conditions the queen introduced some improvements (notably the Urbarial Patent in 1767, which attempted to standardize peasant holdings and obligations), followed their masters in aspiring to nothing more than as much material comfort as could be obtained with a minimum of effort. The national language itself was becoming little more than a peasant dialect, since the language of public administration and the Diet was Latin and of business life was German; and, like the language, the national spirit seemed near moribund.

Joseph II and Leopold II. The nation was shocked out of its lethargy by the accession of Maria Theresa's son, Joseph II, on her death in 1780. Evading the obligation of a king on coronation to swear allegiance to the constitution by not submitting himself to coronation at all (he had the holy crown conveyed to Vienna), Joseph drew Hungary into the Habsburg realm. The counties were transformed into local branches of the state service, taking all their orders from above. German was made the language of government and all education above the elementary level. (A secularized school system had been introduced in 1777.) The land was surveyed in preparation for taxing all estates in it equally. That the position of the peasants was improved pleased them but not their lords. When Joseph fell mortally sick, the country was on the brink of open revolt; on his deathbed he retracted his administrative reforms, but his successor, Leopold II (1790–92), was obliged to restore the ancient constitution and to swear to treat Hungary as a wholly independent kingdom, to be ruled only in accordance with its own laws and customs.

Francis I: the reform generation. When Leopold died in 1792, his son, Francis, made the motions of conforming with his coronation oath before slipping back into the old ways. The Diet was convoked simply to supply money and, after 1811, did not convene for 13 years. Social reaction accompanied this political absolutism, and the stranglehold on economic development was not relaxed.

For many years the Diet, composed either of magnates who identified their interests with those of the court or of landowners who had prospered during the Napoleonic Wars, was as nonprogressive as Francis himself, but in wider circles the spirit of the age had given birth to a great cultural revival that was now bringing forth its first literary fruits, and the new national pride that it embodied was demanding fulfillment of Leopold's promises and an end to the veiled but oppressive dictatorship of Vienna. A great reform movement was set in motion by Gróf (count) István Széchenyi, who proclaimed that the ancient privileges of the nobility were no bastion but a prison. He argued that the servile state of the peasants was degrading and a source of weakness for the nation and also that the system of forced field labour, as well as the nobles' exemption from taxation, was economically harmful even to its supposed beneficiaries. After financial stringency had forced Francis to reconvoke the Diet in 1825, and thereafter to convoke it regularly, doctrines like these were taken up by a whole "reform generation," the most prominent figures of which were the legal expert Ferenc Deák; József, Báró (baron) Eötvös, leader of a small group that opposed breaking with the ruling dynasty; and above all Lajos Kossuth, who largely changed the current of the reform movement by his insistence that social and economic reform could be fully realized only after the achievement of political independence. After Francis had been followed on the throne in 1835 by the luckless Ferdinand—in practice by the government of the two principal ministers, Klemens von Metternich and Anton von Kolowrat—Vienna was driven increasingly onto the defensive and forced to make repeated concessions, especially with respect to the replacement of Latin and German by Magyar as the language of the Diet, administration, and education.

Leaders of the reform movement

The nationalities. This raised a new and painful issue. The population of Hungary, even excluding Croatia, had never been purely Magyar, but the pre-Magyar inhabitants of the plains and the newcomers to them (outside the towns) had quickly become Magyarized; and, while this was not true of the peripheral areas, their populations were relatively sparse. By the end of the 15th century the Slovaks and Ruthenes of the north; the Germans of the free boroughs, the Szepes, and Transylvania; and the Romanians numbered hardly more than 20 to 25 percent of the total. The Magyar majority included almost the entire politically active "noble" class, the non-Magyar recruits to which assimilated most readily. The surviving non-Magyar peasants had neither the wish nor the ability to question the Magyar character of the state, which for its part was uninterested in what languages were spoken by the politically disregarded, unfree populace.

Between 1500 and 1800, however, the ethnic composition of the country changed. The most purely Magyar areas were heavily depopulated during the Turkish wars. These losses were accompanied by mass immigrations of Serbs, Croats, and Romanians from the Balkans and later by the introduction by the Austrian government of large numbers of German and other colonists. By 1720 the Magyars numbered only some 35 percent of the total population. By 1780 the figure had risen to nearly 40 percent, but the periphery was still largely non-Magyar. Moreover, as a result of this ethnic colonization, the population of Hungary grew to nine million by the end of the 18th century, more than double the country's population in 1720.

In this environment the ideas of the French Revolution and of nationalism, one of its major consequences, took hold. Hungarians and most of the other ethnic groups discovered their own national identities. From the late 18th century, poetry, drama, fiction, and literary criticism combined to elevate the Hungarian vernacular to the standard of a literary language, partly in response to the forced Germanization by the Habsburgs but even more as part of an international trend that was particularly strong in central Europe. Institutions such as the national library, the national opera, and the Hungarian Academy of Sciences were also part of the linguistic-cultural movement that soon took the form of self-conscious chauvinism and then became an organized political movement.

Revolution, reaction, and "compromise." The Hungarian reformers' opportunity came in the spring of 1848. Inspired by the Revolution of 1848 in Paris, a popular upheaval caused the breakdown of central authority in Vienna. On March 15 (a date celebrated in Hungary ever since), a bloodless revolution led by young intellectuals, including the poet Sándor Petőfi, abolished censorship in Pest and formulated a series of demands. Seizing the moment, Kossuth prodded the Diet to rush through a body of laws. The April Laws enacted important internal reforms, such as the generalizing of taxes, the abolition of villein status and the transfer of villein holdings to their cultivators, and the reorganizing of the lower "table" of parliament on a representative basis. They also provided for the restoration of the territorial integrity of the lands of the Hungarian crown (subject, in the case of Transylvania, to the agreement of its Diet) and the appointment of a "responsible independent Hungarian Ministry," which was headed by a progressive magnate, Gróf Lajos Batthyány, and included Kossuth, Széchenyi, Deák, and Eötvös. But the new government had enemies: the conservatives resented the land reform, and the centralists (*i.e.,* those who advocated a Vienna-dominated empire) regarded the independent ministry, particularly its three "common" portfolios, as dangerous to the integrity of the monarchy. They found allies among the disaffected nationalities, notably the Serbs and Romanians, and in the Croats, whose ban, Josip Jelačić, refused to recognize the authority of Buda and Pest. Tension between Vienna and Pest mounted steadily, and in September, when the rest of the monarchy had been reduced, Jelačić, on Vienna's orders, invaded Hungary. Batthyány and other ministers resigned, leaving Kossuth in charge. An improvised national army drove Jelačić out of the country, but in December Ferdinand (whose coronation oath bound him to observe the April Laws) was made to abdicate in favour of his young nephew, Francis Joseph. The invasion was now renewed. A panmonarchic constitution abolished the April Laws, in reply to which a rump Diet, inspired by Kossuth, proclaimed the full independence of Hungary and the deposition of the Habsburg dynasty (April 14, 1849). The Hungarian forces, led by a young soldier of genius, Artúr Görgey, held their own until the Austrian court appealed for help to the Russian tsar, who sent an army across the Carpathians. Bitter fighting went on for some weeks more, but the odds were too heavy. On August 12, Kossuth fled the country, transferring his authority to Görgey, who the next day surrendered at Világos to the Russian commander.

Revolution of 1848

Savage reprisals followed, and the country was again subjected to an absolutist and extortionate rule exercised from Vienna through a foreign bureaucracy. This "Bach regime" (for Alexander Bach, Austrian minister of the interior) was maintained until Austria's defeat in Italy in 1859 forced Francis Joseph to begin his retreat from absolutism. The followers of the exiled Kossuth were irreconcilable, but many inside Hungary rallied behind Deák. He held that the April Laws were legally valid and that Hungary's right to complete internal independence was inalienable but that under the Pragmatic Sanction, which he accepted, foreign affairs and defense were subjects "common" to the two halves of the monarchy and that a mechanism could be devised for handling these affairs constitutionally. A Diet convoked in 1861 was dissolved after a few weeks, because the gap between the Hungarians' views and those of Francis Joseph and his centralist ministry in Vienna were still too wide to be bridged. Absolutism was reimposed, but the pressure of international and internal economic difficulties gradually drove Francis Joseph to further concessions. In July 1865 he dismissed his centralist ministry; in December a new Diet was convoked and the negotiations reopened. Interrupted by the outbreak of the Seven Weeks' War, they were resumed after Austria's defeat by Prussia in 1866 had further convinced both parties of the necessity of agreement.

Reprisals from Vienna

The Dual Monarchy, 1867–1918. A new Transylvanian Diet had already approved reunion with Hungary. In February 1867 Francis Joseph, having admitted the validity of the April Laws conditionally on the revision of those dealing with "common" (*i.e.,* overlapping) affairs,

appointed a responsible Hungarian ministry under Gróf Gyula Andrássy. A committee of the Diet then elaborated a law that, while laying down Hungary's full internal independence, provided for common ministries for foreign affairs and defense, each under a joint minister. A third common minister was in charge of the finance for these portfolios. The respective quotas to be paid for these services by each half of the monarchy were reconsidered every 10 years, as were commercial and customs agreements. At first, the two countries formed a customs union. On June 8 Francis Joseph was crowned king of Hungary, and on July 28 he gave his assent to the law.

Treatment of minorities

Francis Joseph had stipulated that the settlement should include a revised Hungaro-Croatian agreement and provisions guaranteeing adequate rights for the non-Magyars of Hungary. The Croatian settlement (1868) left Croatia, including Slavonia, part of the Hungarian crown, under a ban appointed on the proposal of the Hungarian prime minister. Croatia was to enjoy full internal autonomy, but certain questions were designated as common to Croatia and Hungary. When these were under discussion, Croatian deputies attended the central Parliament, in which they could speak in Serbo-Croatian, the sole language of internal official usage in Croatia.

The Nationalities Law (1868) guaranteed that all citizens of Hungary, whatever their nationality, constituted politically "the indivisible, unitary Hungarian nation," and there could be no differentiation between them except in respect of the official usage of languages, and then only insofar as necessitated by practical considerations. The language of the central administrative and judicial services, and of the university, was Hungarian, but there were to be adequate provisions for the use of non-Hungarian languages on lower levels. The consolidation was completed by the incorporation of the Military Frontier and of Transylvania, the latter process involving the abolition of the old "Three Nations," except that the Saxon "university" was allowed to survive as a purely cultural institution.

Hungary under dualism. The Compromise (Ausgleich) of 1867 restored territorial integrity to Hungary and gave it more real internal independence than it had enjoyed since 1526; the monarch's powers in internal affairs were strictly limited. In the conduct of foreign affairs or defense, however, Hungary still formed only part of the monarchy, and its interests in these fields had to be coordinated with those of its other components. But Hungary had a large voice in the monarchy's policy in these fields and enjoyed the great advantage that the resources of the great power of which it formed a part stood behind the country. To some, however, the price still seemed too high, and the parliamentary life of Hungary from 1867 to 1918 was dominated by the conflict between the supporters and the opponents of the Compromise, the latter ranging from complete separatists to those who accepted the Compromise in theory but wanted details of it altered.

The supporters of the Compromise, then known as the Deák Party, held office first but soon got into such financial and personal difficulties that complete chaos threatened. It was averted when in 1875 Kálmán Tisza, the leader of the moderate nationalist Left Centre, merged his party with the remnants of the Deákists on a program that amounted to putting his party's main demands into cold storage until the political and financial situation was stabilized. This new Liberal Party then held office for nearly 30 years. During these years the Compromise stood intact, but there was mounting friction with Vienna over the army, which the Hungarians regarded, with some reason, as imbued with a spirit hostile to themselves; over the economic provisions of the Compromise; and over the question of Hungarian participation in control of the National Bank. An army question in 1889 marked something of a turning point, after which relations between the supporters of the Compromise, behind whom stood the crown, and its nationalist opponents were permanently strained. The tension reached a climax in 1903, when the obstruction of the "national opposition" made parliamentary government practically impossible. The premier, Gróf István Tisza (Kálmán Tisza's son), dissolved Parliament. Elections in January 1905 gave a coalition of "national" parties a parliamentary majority, but Francis Joseph refused to entrust the government to them on the basis of their program, which included "national" concessions over the army. A period of nonparliamentary government followed until April 1906, when the coalition leaders, under threat of an extension of the suffrage if they proved recalcitrant, gave the king a secret undertaking that, if appointed, they would not press the essentials of their program. On this basis he appointed a coalition government, but under a Liberal, Sándor Wekerle. With their hands thus tied, the coalition made a wretched showing. Tisza reorganized the Liberal Party as the Party of National Work, and in the elections of 1910 this party secured a large majority. After Gróf Károly Khuen-Héderváry (1910–12) and László Lukács (1912–13), Tisza himself again became prime minister, and Francis Joseph ceased to press his demand for effective franchise reform, to which Tisza was inexorably opposed.

The rise of the Liberal Party

Social and economic developments. Hungary underwent much change after 1867. The achievements of the Deákist and Liberal governments included the assimilation of the former outlying areas of Transylvania and the Military Frontier, a reform of the relations between the central government and the counties, and a general reorganization of the administration. The judicial system was modernized. Relations between the state and the churches were, after a long struggle, restated in 1894–95 on terms satisfactory to the liberal philosophy of the day, which brought with them full emancipation for Hungary's large Jewish population. In 1868 Eötvös carried through an admirable elementary education act, and much headway was made in raising the educational and cultural level of the country. After long difficulties the national finances were put in order and the public debt reduced. There was considerable economic progress in many fields. Agriculture remained the mainstay of the economy. The medium and small landowners had been hard hit by the land reform of 1848, but the survivors were helped by the high agricultural prices and the secure Austrian market. Afterward, the general European agricultural depression plunged even the big landowners into difficulties, but these diminished near the end of the century, when prices rose again, while the quality and quantity of production improved. Many branches of industry failed to survive the customs union with Austria, but agriculture prospered, and later, as domestic capital accumulated, a process of industrialization, helped by state legislation, set in and expanded rapidly after 1890. As late as 1910 agriculture was still the most important branch of the economy, and 70 percent of the population still derived its livelihood from the soil, while about one-sixth did so from industry and mining.

Economic progress

Urbanization proceeded apace. The growth of Budapest (as it was called from the early 1870s) was meteoric—its population before World War I topped 800,000—and two other cities had populations of more than 75,000 and a dozen around 50,000. Communications were largely modernized.

For all this, Hungary was still a relatively poor country. The extremely rapid growth of the population (from 13 million in 1850 to more than 20 million in 1910) had far outstripped that of the means of production. The growth of industry was still too slow to absorb the surplus rural population, and in spite of high emigration acute rural congestion had developed. While 35 percent of the land was held in 4,000 large estates, there were about 2 million small, or dwarf, holdings, and a further 1.7 million persons (wage earners) were totally landless. A large proportion of these rural workers were forced to live in conditions of extreme misery and near starvation. The living standards and conditions of the industrial workers, especially the unskilled, were also very low.

The political structure was not modern. The unreformed franchise excluded the masses from political influence, and even the vocational organization that they were able to achieve was primitive. The industrial and financial development had been largely the work of Jews (who also held a large part in the professions) or of Magyarized Germans. Its own quasi-alien character and its small numbers prevented the Hungarian middle class from developing

into a positive factor in the political life, which continued to be dominated by a landowning class whose social and political ideas failed to move with the times.

The "nationalities problem" remained intractable. After 1868 Hungarian political philosophy insisted more strongly than ever that the Hungarian state must be Magyar in spirit, in its institutions, and, as far as possible, linguistically. Suggestions to the contrary, or appeals to the Nationalities Law, met with derision or abuse. In spite of the law, the use of minority languages was banished almost entirely from administration and even justice. While the autonomy of the church schools was hardly attacked until the 20th century, most denominations saw to it that all secondary education in their schools, with trivial exceptions, was in Hungarian, which was also overrepresented in the primary schools, as it was in practically all instruction in the state schools founded from 1870 onward.

By the end of the century, the state apparatus was entirely Hungarian in language, as were business and social life above the lowest levels. The proportion of the population with Hungarian as its mother tongue rose from 46.6 percent in 1880 to 51.4 percent in 1900. The Magyarization of the towns had proceeded at an astounding rate. Nearly all middle-class Jews and Germans and many middle-class Slovaks and Ruthenes had Magyarized.

Magyarization of the middle classes

Most of the Magyarization, however, had been in the centre of Hungary and among the middle classes. It had hardly touched the rural populations of the periphery, and the linguistic frontiers had hardly shifted from the line on which they had been stabilized in the 18th century. In these areas, moreover, a hard core of national feeling had survived. This had weakened during the first decades after the Compromise but was reviving again at the beginning of the 20th century, especially among the Romanians, and was being encouraged from across the frontiers of Romania and Serbia and (in the case of the Slovaks) from Bohemia. Hungaro-Croatian relations, too, deteriorated, after a period of quiescence, when the Serbian government began propagating a theory of "Yugoslav" unity designed to detach the Croats from the monarchy.

Many of these developments threatened the very basis of the Compromise, and to this another uncertainty was added. Francis Joseph could be trusted to support and accept the policies of any Hungarian government that on its side maintained the Compromise loyally; but he was an old man, and his heir presumptive, the archduke Francis Ferdinand, was notoriously hostile to the Hungarian regime. In touch with many of its opponents, the archduke was credited with designs of overthrowing the Compromise to the benefit not of its traditional opponents, the Hungarian Independents, but of its enemies in the opposite camps, especially the nationalities.

World War I. The assassination of the archduke on June 28, 1914, removed this danger and plunged Austria-Hungary into World War I. For the first two years of the war, Tisza upheld the internal system and held the country to its international course; and when Francis Joseph died he persuaded the new king, Charles IV, to accept coronation (December 1916), thus binding himself to uphold the integrity and constitution of Hungary. Charles, however, insisted on electoral reform, and Tisza resigned (May 1917). While short-lived minority governments struggled with increasing difficulties, a threefold agitation grew: of Hungarian nationalists, against a war into which, they maintained, Hungary had been drawn in the interest of Germany and Austria; of the political left, growing daily more radical under the stimuli of privation and the Russian Revolution of 1917; and of the nationalities, encouraged by the favour that their kinsfolk were finding with the Triple Entente. The country began to listen to Gróf Mihály Károlyi, leader of a faction of the Independence Party, who proclaimed that a program of independence from Austria, repudiation of the alliance with Germany, and peace with the Entente, combined with social and internal political reform and concessions to the nationalities, would safeguard Hungary against all dangers at once. Hungary's submergence in the long, devastating war included the mobilization of 3.5 million men and exhausted the Hungarian economy. Agricultural output declined by half during the last years of the war, and the currency lost more than half of its value. In the autumn of 1918 Hungary was on the brink of economic collapse.

REVOLUTION, COUNTERREVOLUTION, AND THE REGENCY, 1918–45

On Oct. 31, 1918, when the defeat of the monarchy was imminent, Charles appointed Károlyi prime minister at the head of an improvised administration based on a left-wing National Council. After the monarchy had signed an armistice on November 3 and Charles had "renounced participation" in public affairs on the 13th, the National Council dissolved Parliament on the 16th and proclaimed Hungary an independent republic, with Károlyi as provisional president. The separation from Austria was popular, but all Károlyi's supposed friends disappointed him, and all his premises proved mistaken. Serb, Czech, and Romanian troops installed themselves in two-thirds of the helpless country, and, in the confusion, orderly social reform was impossible. The government steadily moved leftward, and in March 1919 Károlyi's government was replaced by a soviet republic, controlled by Béla Kun, who had promised Hungary Russian support against the Romanians. The help never arrived, and Kun's doctrinaire Bolshevism, resting on a "Red terror," antagonized almost the entire population. On August 4 Kun and his associates fled Budapest, and two days later Romanian troops entered the city. Shadow counterrevolutionary governments had already formed themselves in Szeged and Vienna and pressed the Allies to entrust them with the new government. The Allies insisted on the formation of a provisional regime including democratic elements that would be required to hold elections on a wide, secret suffrage. The Romanians were, with difficulty, induced to retire across the Tisza River, and a government, under the presidency of Károly Huszár, was formed in November 1919. Elections were held in January 1920. The new Parliament declared null and void all measures enacted by the Károlyi and Kun regimes as well as the legislation embodying the Compromise of 1867. The institution of the monarchy was thus restored, but its permanent reinstatement was predicated on the resolution of the differences between the nation and the dynasty, an issue that divided Hungarians. In the interim Admiral Miklós Horthy, who had organized the counterrevolutionary armed forces, was elected regent as provisional head of state (March 1, 1920). The Huszár government then resigned, and on March 14 a coalition government, composed of the two main parties in the Parliament (the Christian National Union and the Smallholders), took office under Sándor Simonyi-Semadam.

The soviet republic of Béla Kun

The Regency, 1920–45. *The Treaty of Trianon.* The Allies had long had their peace terms for Hungary ready but had been unwilling to present them to an earlier regime. It was, thus, the Simonyi-Semadam government that was forced to sign the Treaty of Trianon (June 4, 1920). The Allies not only assumed without question that the country's non-Hungarian populations wished to leave Hungary but also allowed the successor states, especially Czechoslovakia, to annex large areas of ethnic Hungarian population. The final result was to leave Hungary with only 35,893 of the 125,641 square miles that had constituted the lands of the Hungarian crown. Romania, Czechoslovakia, and Yugoslavia took large fragments, while others went to Austria and even Poland and Italy. Of the population of 20,866,447 (1910 census), Hungary was left with 7,615,117. Romania received 5,257,467; Czechoslovakia, 3,517,568; Yugoslavia, 4,131,249; and Austria, 291,618. Of the 10,050,575 persons for whom Hungarian was the mother tongue, no fewer than 3,219,579 were allotted to the successor states: 1,704,851 to Romania, 1,063,020 to Czechoslovakia, 547,735 to Yugoslavia, and 26,183 to Austria. While the homes of some of these—*e.g.,* the Szeklers—had been in the remotest corners of historic Hungary, many were living immediately across the frontiers.

In addition, the treaty required Hungary to pay in reparations an unspecified sum, which was to be "the first charge upon all its assets and revenues," and limited its armed forces to 35,000, to be used exclusively for the maintenance of internal order and frontier defense.

Postwar confusion and reconstruction. Conditions in Hungary in 1920 were exceedingly difficult in every respect. The prolonged war, the Bolshevik regime (before which mobile capital had fled headlong), and the rapacious Romanian occupation had exhausted its resources, and the economy had been further disrupted by the new frontiers, which cut factories off from both their accustomed supply sources and their markets. Industrial unemployment had reached unprecedented heights, and the surviving national resources were being strained to support nearly 400,000 refugees from the successor states.

Both industrial and agrarian workers were embittered by the failure of their revolutionary hopes. Even more dangerous were the armies of the "new poor," not only the homeless refugees but also a large part of the middle classes in general, reduced to penury by the galloping inflation.

The right-wing radicals

They formed a radical army, one of the right that ascribed their misery precisely to the revolutions, on which they put the blame for all Hungary's misfortunes. Feelings ran particularly high against the Jews, who had played a disproportionately large part in both revolutions, especially Kun's, but the resentment extended also to the Social Democrats and even to Liberal democracy. "White terrorists" wreaked indiscriminate vengeance on persons whom they associated with the revolutions. Huszár's government itself had turned so sharply on the Social Democrats and the trade unions that the former withdrew their representatives from the government and boycotted the elections, in protest against the widespread arrests and internments (roughly 5,000 people were executed and 70,000 interned or imprisoned). Communists, radical democrats, and Jewish intellectuals emigrated in large numbers.

The government of Pál Teleki, who succeeded Simonyi-Semadam in July 1920, blunted the edge of the agrarian unrest with a modest reform—promised as a first installment—that took 1.7 million acres (7.5 percent of the total area of the country) from the biggest estates for distribution in smallholdings, but it had hardly touched any other social problem when in March 1921 the legitimist question was raised by King Charles's sudden return to Hungary. He was ordered to withdraw by the Allies with the willing compliance of the right-wing radicals, toward whom Horthy was then leaning. The government, several of whose members were legitimists, resigned, and the succession was assumed by the conservative Gróf István Bethlen, who had been waiting behind the scenes. Bethlen devised a formula that, while not legally excluding the king's return (under Entente pressure, Parliament had voted a law dethroning the Habsburgs, but even Hungary's own anti-legitimists never took it as morally binding), excluded it in practice. In return for this, the Smallholders Party agreed with the antilegitimists among the Christian nationalists to form a new Party of Unity under Bethlen's leadership. He persuaded Parliament to accept as still legally in force the franchise enacted in 1918, which reduced the number of voters and reintroduced open voting in rural districts. Elections in May 1922 gave this party a large majority.

The end of legitimism

Meanwhile, a second attempt by the king (in October 1921) to recover his throne again failed, and, soon after, the legitimist question lost its acuteness with Charles's death. In December 1921 Bethlen concluded a secret pact with the Social Democrats, under which the latter promised to abstain from political agitation and to support the government's foreign policy in return for the end of persecution, the release of political prisoners, and the restoration of the sequestrated trade union funds. The peasant leaders were persuaded to accept the indefinite postponement of further land reform. The "White terror" was liquidated quietly but effectively, chiefly by finding government employment for the right-wing radical leaders.

Bethlen's domestic program was made possible by his cautious international policy. Almost all Hungarians were passionately convinced of the injustice of the Treaty of Trianon, the redress of which was the all-dominant motive of Hungary's foreign policy throughout the interwar period and the key to the hostile relations between Hungary and those states that had chiefly profited by it. Bethlen was as revisionist at heart as any of his countrymen, but he was convinced that Hungary could not act effectively in this field until it had acquired friends abroad and had achieved political and economic consolidation at home. This depended on financial reconstruction; to achieve this, he applied for Hungary's admission to the League of Nations, which was granted (not without difficulty) in September 1922. In March 1924, in return for an agreement to carry out loyally the obligations of the treaty, he obtained a League loan, which had almost magical effects. Inflation stopped immediately. The League loan was followed by a flood of private lending, and the expatriated domestic capital returned. With this help, Hungary enjoyed some years of prosperity, during which agriculture revived and industrialization made progress.

Abroad, Bethlen's only other important move was the conclusion in 1927 of a treaty of friendship with Italy. At home his regime, which was conservative but not tyrannical, rested on Hungary's conservative-liberal forces, to the exclusion of extremism from left or right.

Financial crisis: the rise of right radicalism. Bethlen's command of Parliament was complete and unshaken by the disastrous fall in world wheat prices in 1929. In June 1931 he had just held elections that returned his party with its usual large majority when a world financial crisis supervened on the economic one to shatter the foundations of his structure. Foreign creditors called in their money, and Hungary, its trade balance annihilated by the collapse of the wheat market, could not meet their demands and had to apply for help from the League of Nations, which imposed a regime of rigid orthodox deflation. Industrial unemployment soared again, the agricultural population was rendered almost literally penniless, and the government services had to carry through large-scale dismissals and salary reductions in the interests of a balanced budget.

In August Bethlen resigned. His successor, Gróf Gyula Károlyi, was unable to cope with the situation. Political agitation mounted, and on Oct. 1, 1932, Horthy appointed as prime minister the leader of the right-wing radicals, Gyula Gömbös.

Gömbös appointed prime minister

At home Gömbös found the leading strings of financial forces, international and domestic, as invincible as had his predecessors. Previously a violent anti-Semite, he had to recant his views on this point and was unable to carry through any other points of his fascist program, particularly as Horthy at first refused to allow him to hold elections. Neither was he able to realize his foreign political ideal of an "Axis" composed of Hungary, Italy, and Germany, since his two proposed partners were then at loggerheads over Austria, and Gömbös, one of whose first acts had been to dash to Rome and breathe new life into Hungary's friendship with Italy, found himself drawn into the "Rome Triangle" (Italy, Austria, and Hungary) that was directed precisely against Germany. Finally, Adolf Hitler upset another of his calculations by telling him that, while Germany would help Hungary against Czechoslovakia, it would not do so against Romania or Yugoslavia.

Nevertheless, by the time of Gömbös' premature death in October 1936, he had managed to achieve at least some of his goals. Shortly before Gömbös died, Horthy had allowed him to hold elections, which brought into Parliament a strong right-wing radical contingent from which it could never thereafter free itself. Abroad, when Benito Mussolini became subordinate to Hitler, Hungary found itself after all in a sort of Axis camp, membership of which might help it at least to accomplish partial revision of the Treaty of Trianon. On the other hand, if Germany chose to apply economic or political pressure, Hungary would be defenseless but for such shadow help as Italy could offer.

This threat already loomed large, and it became inextricably involved with Hungary's internal politics, by reason of the ideological character of the Nazi regime and in particular its anti-Semitism. Anti-Semitism at that stage was running high in Hungary itself, and those infected by it—not just the right-wing radicals of various brands but other members of the middle classes as well—welcomed Germany's support for their own ideas while making light of its dangers. They even argued, not without reason, that the danger lay in affronting Germany, which could easily crush unarmed little Hungary but would not wish to attack a friend and ideological partner. They further believed

The question of support for Germany

(most army officers held this view) that, should Hitler's policies lead to war, Germany would emerge the victor; Hungary's salvation thus lay in joining forces with Germany. On the other side a curious shadow front emerged, composed of all elements antagonistic to Nazism—not only Hungary's Jews but also the legitimists, the traditionalist conservative-liberals, and the Social Democrats. Many of these were not convinced that Germany was invincible and held that, if war came, only disaster could follow for Hungary if it became too closely involved with Germany. Even they, however, were unwilling to draw the ultimate conclusion that Hungary should abandon all its revisionist claims and join hands with the Little Entente, which, for its part, indicated that it would accept nothing short of total renunciation. It was of the highest importance that by this time Horthy had shed his earlier right-wing radical leanings and sympathized with this shadow front.

To succeed Gömbös, Horthy appointed Kálmán Darányi, more of a conservative than a right-wing radical; the appointment was ill-received in Germany, which grew more hostile the next year, when Darányi's foreign minister, Kálmán Kánya, obtained the tacit consent of the Little Entente for Hungary to rearm, although Hungary was still short of armaments, for which Germany was its only source of supply. On a visit to Berlin, Darányi and Kánya smoothed over the difficulties; but, when Darányi tried to placate the extremists at home, Horthy replaced him (in May 1938) with Béla Imrédy, who introduced a largely token "Jewish Law" but nevertheless pinned his hopes on the West. When the Munich crisis broke in September, Imrédy and Kánya, while presenting Hungary's claims on Czechoslovakia, limited the claims to what they hoped would be acceptable to the Western powers, whose endorsement they made every effort to obtain. Ignored by the West, the Hungarian leaders had to turn to Germany and Italy, which, under the "First Vienna Award" of November 2, gave Hungary the fringe of southern Slovakia inhabited by ethnic Hungarians. Imrédy, disillusioned with the West, dismissed Kánya for the pro-Axis Gróf István Csáky and sought to recover Hitler's favour by introducing a more far-reaching Jewish Law. Imrédy's enemies secured his resignation in February 1939 by unearthing documents purporting to show a Jewish strain in his own ancestry. Pál Teleki, who succeeded him, was a most convinced Westerner, but Hungary's recovery of Carpatho-Ruthenia (March 1939) was, again, sanctioned by Hitler.

War and renewed defeat. When Germany attacked Poland (Sept. 1, 1939), Hungary refused to allow German troops to cross Hungarian territory. In the first months of World War II none of the belligerents wanted the war to extend to southeastern Europe, so Teleki and Horthy were able to keep Hungary at peace. After the Soviet Union had occupied Bessarabia in June 1940, the Hungarian leaders compelled a reluctant Germany (but a willing Italy) to cede to Hungary northern Transylvania under the "Second Vienna Award." But they then allowed German troops to cross Hungarian territory into southern Romania and in November signed the Tripartite Pact.

German troops in Hungary

The next step was more fatal still. In his search for cautious reinsurance, Teleki concluded with the like-minded government of Yugoslavia a treaty unluckily characterized as one of "Eternal Friendship." On March 26, 1941, that Yugoslav government was overthrown by a pro-Western regime. Hitler prepared to invade Yugoslavia and called on Hungary to help him. Hungary refused to join in the attack but again allowed German troops to cross its territory. Great Britain threatened to declare war, and Teleki, blaming himself for the development of a situation that it had been his life's aim to avoid, took his own life on April 2. His successor, László Bárdossy, waited until Croatia had declared its independence (April 10); then, arguing that Yugoslavia had already disintegrated, he occupied the ex-Hungarian areas of Yugoslavia.

Although he was not a fascist, Bárdossy believed that the Axis powers would win the war and that Hungary's salvation lay in placating them; otherwise, he believed, Romania (now pro-Axis) would persuade Hitler to reverse the Second Vienna Award. Accordingly, when Germany attacked the Soviet Union (June 22), Bárdossy sent a token force to assist in what everyone expected to be a brief operation. The strength of the Soviet resistance upset the calculation, and in January 1942 the Germans forced Hungary to mobilize practically all its available manpower and send it to the Soviet Union. Meanwhile, Britain, by this point allied with the Soviet Union, declared war (December 1941) on Hungary, which in its turn declared war on the United States. Further, Britain recognized the Czechoslovak government-in-exile and withdrew recognition of the First Vienna Award, and the Soviet Union recognized Czechoslovakia's 1937 frontiers.

Hungary in the war

Many Hungarians by then agreed with Bárdossy that Hungary's only course was to fight on until the Axis won the war—the more so because all Hungarians except those of the extreme left regarded Bolshevism as the embodiment of evil. Horthy, however, while sharing the latter view, still believed in a Western victory and thought it possible for Hungary, while continuing the struggle in the east, to regain the favour of the West. In March 1942 he replaced Bárdossy with Miklós Kállay, who shared these hopes. For two years Kállay conducted a remarkable balancing act—protecting Hungary's Jews and allowing the left (except for the communists) almost untrammeled freedom while putting out innumerable feelers to the Western Allies, to which he actually promised to surrender unconditionally when their troops reached Hungary's frontiers. Meanwhile, in January 1943 the Hungarian expeditionary force suffered a crushing defeat at Voronezh in western Russia that cost it much of its manpower and nearly all its equipment. (C.A.M./Ed.)

But the Western forces did not approach the Danube valley, and, as the Soviet army neared the Carpathians, Hitler decided that he could not leave his vital communications at the mercy of an untrustworthy regime. In March 1944 he offered Horthy the choice between full cooperation under German supervision or undisguised German occupation with the treatment accorded to an enemy country. Horthy chose the former course and appointed a collaborationist government under Döme Sztójay. For a while the Germans did much as they wished—they suppressed parties and organizations of potential opponents and arrested their leaders. With the cooperation of Hungarian authorities, Jews were compelled to wear a yellow star, robbed of their property, and incarcerated in ghettos as in other Nazi-occupied areas. Except for the Jews in the capital and those in the forced-labour camps of the Hungarian army—whose turn would come later—they were deported to the gas chambers of German concentration camps. In spite of the efforts of representatives of some neutral countries—such as Raoul Wallenberg of Sweden, the papal nuncio, and diplomats of Switzerland, Portugal, and even Spain, who saved tens of thousands of lives—more than 550,000 of Hungary's nearly 750,000 Jews, as defined by racist legislation, perished during the war.

In the summer the pressure relaxed; and in August, after Romania's surrender to the Allies, Horthy appointed a new government under the loyal general Géza Lakatos and again extended peace feelers. A "preliminary armistice" was concluded in Moscow, but, when on October 15 Horthy announced this on the radio, he was abducted by the Germans, who forced him to recant and to abdicate. The Germans put Ferenc Szálasi, the leader of the right-wing extremist Arrow Cross movement, in charge. By then, however, Soviet troops were far inside the country. The Germans and their Hungarian allies were driven back slowly, while numerous refugees fled with them. The last armed forces crossed the Austrian frontier in April 1945.

Soviet troops in Hungary

The defeat was sealed in a new peace treaty, signed in Paris on Feb. 10, 1947, which restored the Trianon frontiers, with a rectification in favour of Czechoslovakia and the Soviet Union. It imposed on Hungary a reparations bill of $300 million and limited its armed forces. The implementation of the treaty's provisions was to be supervised by a Soviet occupation force, a large contingent of which remained in the country.

HUNGARY SINCE 1945

As in 1920, a new regime recognized the defeat of its predecessor. As early as December 1944 a makeshift Pro-

visional National Assembly had accepted a government list and program presented to it by communist agents following in the wake of the Soviet armies. Beginning cautiously, the communists announced that the new Hungary was to rest on "all its democratic elements." The government contained only two communists; its other members were representatives of four noncommunist left-wing parties—the Smallholders, the Social Democrats, the National Peasants, and the Progressive Bourgeoisie—and four men associated with the Horthy regime. The program provided for the expropriation of the large estates and the nationalization of the banks and heavy industry; but it promised guarantees of democratic rights and liberties, respect for private property, and encouragement of private initiative in trade and small industry.

The communist regime. *Political developments.* The full political takeover, however, proceeded systematically, although not according to any timetable, for the communists, misjudging feeling in the country, allowed the first elections (November 1945) to be relatively free. Only the parties of the coalition were allowed to contest them; but the adherents of the proscribed parties voted for the Smallholders, who received an absolute majority. The head of the Soviet mission, however, insisted that the coalition must be maintained; a Smallholder was allowed to be prime minister, but the Ministry of the Interior, with the control of the police, was given to the communists. Pressure and intimidation were then applied to the Smallholders to expel their more courageous members as "fascists"; and in the next election (August 1947) the Smallholders polled only 15 percent of the votes cast. The communists had meanwhile forced the Social Democrats to form a "workers' bloc" with them. Although the pressure this time was considerable, the bloc still polled only 45 percent of the votes (other parties were allowed to participate this time); however, the communists then forced the Social Democrats to join them in a single Workers' Party, from which recalcitrants were expelled.

Communist takeover

In the next election (May 1949) voting was open, and the voters were presented with a single list, on which candidates identified as Smallholders and National Peasants were actually tools of the communists. In August a new constitution was enacted—a copy of that of the Soviet Union. Hungary, a republic since Feb. 1, 1946, now became a "people's republic," and, although its president (Zoltán Tildy) and for a while its prime ministers (Ferenc Nagy, then Lajos Dinnyés) were Smallholders, all real power rested with the Workers' Party, which was controlled by its first secretary, Mátyás Rákosi. Finally, the party's "Muscovite" wing turned on its "national" wing. The leader of this latter group, László Rajk, was executed on questionable charges in October 1949, and his chief adherents were similarly executed or imprisoned. Meanwhile, hundreds were executed or imprisoned as war criminals, many of them for no other offense than loyalty to the Horthy regime. Many thousands more were interned. The State Security Department (AVO) was omnipotent. The judiciary, civil service, and army were purged, and party orthodoxy became the criterion for positions in them. The trade unions were made into mere executants of party orders. After the dissolution of the parties, the chief ideological opposition to the communist regime came from the churches; but their estates were expropriated, making it impossible for them to maintain their schools, and in 1948 the entire educational system was nationalized. The Calvinist and Lutheran churches accepted financial arrangements imposed by the state. The head of the Roman Catholic church, József Cardinal Mindszenty, who refused to follow their example, was arrested in December 1948 and condemned to life imprisonment. The monastic orders were dissolved. Thereafter, the Roman Catholic church accepted financial terms similar to those offered to the other churches, and eventually the bishops, with visible repugnance, took the oath of loyalty to the state.

Economic developments. The communists' economic program, like their political program, could not be realized immediately, because in 1945 Hungary was in a state of economic chaos worse even than that of 1918. This time the country had been a theatre of war. Many cities were in ruins, and communications were wrecked; the retreating Germans had destroyed the bridges between Buda and Pest and had taken with them all they could of the country's portable wealth. The Soviet armies lived off the land, and the Soviet Union took its share of reparations in kind, placing its own values on the objects seized. It also took over former German assets in Hungary, including Jewish property confiscated during the Nazi occupation.

A three-year plan introduced in August 1947 was devoted chiefly to the repair of immediate damage. This was declared completed, ahead of schedule, in December 1949. By then the communists were in full political control, and measures nationalizing banking, most industry, and most internal and all foreign trade had been enacted. Hungary joined other Soviet-bloc countries in founding the Council for Mutual Economic Assistance (Comecon) in 1949. The land, outside the big estates, was not touched at first, but in 1948 Rákosi announced a policy of collectivization of agriculture.

Economic programs

The three-year plan was succeeded by a five-year plan, the aim of which was to turn Hungary into a predominantly industrial country, with an emphasis on heavy industry. Huge sums were devoted to the construction of foundries and factories, many of them planned with little regard for Hungary's real resources and less still for its needs. In fact, the plan was concerned with the needs of the Soviet Union, for which Hungary was to serve as a workshop. Hungary's newly discovered deposits of uranium went straight out of the country. Industrial production rose steeply, but the standard of living did not; the production of consumer goods was throttled and that of agriculture stagnated.

The revolution of 1956. Rákosi—who in 1952 came to preside over the government as well as the party—was all-powerful until the death of Stalin in 1953, when a period of fluctuation began. In July 1953 Rákosi was deposed from the prime ministership in favour of Imre Nagy—a "Muscovite" but a Hungarian in his attitudes and not unpopular in the country. Nagy promised an end to the forced development of heavy industry, more consumer goods, no more forcing of peasants into the collectives, the release of political prisoners, and the closing of internment camps. He introduced some of these reforms, but Moscow hesitated to support him. In the spring of 1955 Nagy was dismissed from office and expelled from the party. Rákosi, reinstated, put the country back on its previous course but was dismissed again in July 1956, this time in disgrace. The new Soviet leader, Nikita S. Khrushchev, had sacrificed Rákosi as a gesture to the Yugoslavian leader Josip Broz Tito, whom he wished to placate and whom Rákosi had offended personally. The new leader, Ernő Gerő, Rákosi's deputy, was almost as detested as Rákosi himself. Gerő promptly announced that there would be no concessions on matters of principle to Nagy and his group.

The relaxation of pressure under Nagy (though transitory), Khrushchev's "secret speech" denouncing Stalin's cult of personality delivered at the 20th Congress of the Communist Party of the Soviet Union (February 1956), and the Polish challenge to the Soviet Union in the spring and summer of 1956 emboldened Hungarians. On October 23, students in Budapest staged a great procession, which was to end with the presentation of a petition asking for redress of the nation's grievances. People flocked into the streets to join them. Gerő answered with an unwise and truculent speech, and police fired into the crowds. The shots turned a peaceful demonstration into a revolutionary one. The army joined the revolutionaries, and army depots and munitions factories handed out arms. Outside Budapest local councils sprang up in every centre; the peasants reoccupied their confiscated fields. The communist bureaucracy melted away. Prison doors were opened. The members of the AVO fled if they could. A cheering crowd escorted Cardinal Mindszenty back to the palace.

Revolution and change

In kaleidoscopic political changes, Nagy resumed power but was driven from one concession to the next, until he found himself at the head of a genuine coalition government composed of Smallholders, Social Democrats, and National Peasants, which, with a "Catholic Association," had reconstituted themselves.

The Soviet troops had withdrawn, and Nagy was negotiating for the complete evacuation of Hungary. On November 1 he announced Hungary's withdrawal from the Warsaw Pact and asked the United Nations to recognize Hungary as a neutral state. Soviet officials were uncertain whether to let matters take their course, but Nagy's denunciation of the Warsaw Pact was too threatening, and their tanks, which had halted just across the frontier, began to return, reinforced by other units. On November 4 the tanks entered Budapest. Nagy took refuge in the Yugoslav embassy and Cardinal Mindszenty in the U.S. legation. General Pál Maléter, head of the Hungarian national forces, was imprisoned.

A communist leader, Ferenc Münnich, and János Kádár, a "National Communist" who had been imprisoned under Rákosi and had actually joined the revolutionaries on October 24, formed a new "revolutionary peasant-worker government," consisting entirely of communists, with Kádár as prime minister. Kádár promised that when the "counterrevolution" had been suppressed and order restored he would negotiate on the withdrawal of the Soviet garrison (although the denunciation of the Warsaw Pact was retracted); he dissociated himself from the "Rákosi-Gerő clique" and promised internal reforms.

The country was not convinced, and fighting broke out. But the odds were too heavy, and the major hostilities were over within a fortnight. The workers, however, proclaimed a general strike, and it was many weeks before they were brought to heel.

The end of the revolution

Meanwhile, Nagy, who had left his place of refuge under safe conduct, had been abducted to Romania. After a secret trial he, Maléter, and a few close associates were executed in 1958. Many lesser figures were seized and transported to the Soviet Union, some never to return; 200,000 refugees escaped to the West. Thus, a substantial proportion of Hungary's educated classes was lost to the country. Material damage was also heavy.

The Kádár regime. In the first uncertain weeks of his regime Kádár made many promises. Workers' councils were to be given a large amount of control in the factories and mines. Compulsory deliveries of farm produce were to be abolished, and no compulsion, direct or indirect, was to be put on the peasants to enter the collectives. The five-year plan was to be revised to permit more production of consumer goods. The exchange rate of the ruble and forint was to be adjusted and the uranium contract revised. For a time there was even talk of a coalition government.

The larger hopes were dashed after representatives of the Soviet Union, East Germany, Czechoslovakia, Romania, and Bulgaria conferred with those of Hungary in Budapest in January 1957. A new program was soon issued stating that Hungary was a dictatorship of the proletariat, which in foreign policy relied on the Soviet Union and the Soviet bloc. Further, it was asserted that the Soviet garrison was in Hungary to protect the nation from imperialist aggression. Internal reforms were again promised, however, and foreign trade agreements were to be based on complete equality and mutual advantage.

Subsequently, Kádár was at great pains to give the Soviet Union no cause for uneasiness over Hungary's loyalty. When any international issue arose, he invariably supported Moscow's policy with meticulous orthodoxy, even sending a contingent into Czechoslovakia in 1968. At home he ignored some of his promises and honoured others only superficially. The peasants were so greatly pressured to enter cooperatives that within a few years practically no private farms survived. The workers' councils were dissolved, but trade unions were later granted rights to query decisions by management. Parliament remained a rubber stamp, and a Patriotic People's Front (PPF), on which noncommunists were represented, was a mere façade.

Improvements under Kádár

Nevertheless, conditions changed very much for the better. Kádár enunciated the principle that "he who is not against us is with us," which meant ordinary people could go about their business without fear of molestation or even much surveillance and could speak, read, and even write with reasonable freedom. Technical competence replaced party orthodoxy as a criterion for posts of responsibility. More scope was allowed to private small-scale enterprise in trade and industry, and the New Economic Mechanism (NEM), initiated in 1968, introduced the profit motive into state-directed enterprises. Agricultural cooperatives were allowed to produce industrial goods for their own use or to sell on demand, while the private plots of their members supplied a large proportion of fruits and vegetables for the rest of the population. Contacts with the West were encouraged. A modus vivendi was found with the Vatican and with Protestant churches. The standard of living began to rise substantially. Tourism developed as a significant industry. In addition to a huge influx of foreign visitors to Hungary, an increasing number of Hungarians traveled abroad, especially after the introduction (Jan. 1, 1988) of "global passports," which removed restrictions on travel. Income from tourism increased dramatically, yet the net balance was less in Hungary's favour than would be expected because Hungarians going to the West spent most of their official hard currency quotas on purchases of consumer goods, owing to shortages and skyrocketing prices at home.

The two decades of the NEM, which went beyond the liberalization that took place in the Soviet Union itself, were only partially successful. Productivity failed to rise according to expectations. Government regulations persisted in many areas, and the economy remained geared to the Soviet-led Comecon. A burdensome system of subventions aimed at keeping down the prices of basic necessities and services and at promoting the production of state-preferred goods made realistic cost accounting impossible. The price rise of petroleum and other industrial raw materials on the world market also aggravated the situation. The gap grew between the price of energy, sophisticated industrial hardware, and raw materials on the one hand and the price of agricultural products, a main item in Hungary's foreign trade, on the other. (C.A.M./G.Ba./Ed.)

Political opposition to reform, including Soviet and Comecon criticism of the NEM, all but brought it to a halt in 1973–78, when administrative interventions by state agencies and party and trade union organizations caused a return to the methods of the centralized command economy under the pretext of protecting the relative earnings of industrial workers compared with those in agriculture or of taxing only "unearned" profits of successful enterprises. Rezső Nyers, the architect of the NEM, was demoted in 1974, only to be brought back to the Politburo in May 1988, at a time of deepening political and economic crisis. By the end of the 1970s, reformers again prevailed. New measures included cuts in the central bureaucracy, encouragement of small firms and private enterprises, revisions of the price and wage system to reflect more closely conditions on the world market and costs of production, and the creation of a commercial banking system.

Reforms of the late 1980s. *Economic reforms.* The efforts to introduce market reforms into Hungary's socialist economy extended to the international arena. Already a member of the General Agreement on Tariffs and Trade (GATT), Hungary was admitted to the International Monetary Fund (IMF) in 1982 and received assistance from the World Bank. An agreement with the European Economic Community (now the European Union), the first among members of Comecon, was also concluded. While the Soviet Union remained Hungary's most important trading partner and the source of its energy supply, Hungary had to turn to the West for technological assistance and capital investment in the process of modernizing the economy. Trade relations with the West, in which Austria and West Germany played particularly important roles, were crucial at a time when barely half of Hungary's foreign trade involved members of Comecon. Foreign trade constituted a larger proportion of Hungary's gross national product (GNP) than that of any other Comecon country.

Economic deterioration

Efforts to adjust Hungary's economy to the world market were handicapped by the adverse effects of the energy crisis of the 1970s and the de facto reversal of the NEM in the same decade. Although agricultural production continued to advance, in part because of favourable international market conditions, the rest of the economy deteriorated. This process was further aggravated by misallocation of

funds, reluctance to abandon costly projects such as the Danube hydroelectric power plant, participation in joint projects of Comecon, and unwillingness to drastically reduce subsidies to inefficient enterprises and for many basic necessities and services, whose price level was kept artificially low for many years. As a result Hungary's hard currency debt reached $18 billion by the end of the 1980s, representing the highest per capita indebtedness of any country in eastern Europe. Inflationary pressures began to build up, and real wages and living standards declined.

The appointment of Károly Grósz as prime minister in mid-1987 led to a program of severe belt-tightening; a harsh, hastily prepared income-tax law aimed at cutting consumption; anticipated unemployment in some segments of the economy; and steep rises in consumer prices, transportation costs, and basic services such as gas, electricity, telephone, water, and rents. Minor changes in the party leadership, still controlled by Kádár, and the reshuffling of the government eased acceptance of unpopular measures introduced to stabilize the collapsing economy.

Political changes and reforms. By the late 1980s, growing numbers of Hungarians had concluded that years of misgovernment could not be erased by economic reforms alone. The process of de-Stalinization reinforced the desire to reexamine the political premises of Grósz's program, which seemed to imply that to keep their hard-won personal freedoms Hungarians should pay with economic misery and further social polarization. By the time the annual inflation rate reached 17 percent, public pressure compelled the party conference in May 1988 not only to replace Kádár with Grósz but also to replace several of Kádár's supporters in the Politburo and Central Committee. Miklós Németh became the prime minister in November 1988; then in 1989 a quadrumvirate of Imre Pozsgay, Grósz, Németh, and Nyers, chaired by the latter, temporarily took over the direction of a deeply split party. In October the party congress announced its transformation into the Hungarian Socialist Party (MSzP). A splinter group of conservatives retained the old name and continued allegiance to its policies.

Weakening of the Politburo

Meanwhile, informal associations, clubs, and debating circles proliferated and served as points of departure for new political parties. The Democratic Union of Scientific Workers, supported by a substantial portion of academic and clerical employees of scholarly institutions, was the first independent professional association to challenge the communist-controlled National Council of Trade Unions and to establish contact with the Polish Solidarity as well as organized labour in the West. Filmmakers, writers, and journalists rediscovered their right of free speech, publishers printed manuscripts that had been kept locked up for decades, new periodicals appeared, and the press, radio, and television threw over taboos that had prevailed for more than 40 years.

The 950th anniversary of the death of Stephen I was celebrated with medieval pomp in August 1988 in the presence of the primate of Poland, Józef Cardinal Glemp, representing Pope John Paul II. Full diplomatic relations with the Vatican were reestablished in 1990. The state also returned to the Protestant churches some of their former prestigious educational institutions, the Boy Scouts (a conspicuously Christian organization in Hungary) was resuscitated, and the World Jewish Congress held its executive session in Budapest in 1987.

The Kádár regime had tried to avoid offending fraternal communist governments by not raising questions about violations of the rights of Hungarian minorities within their jurisdiction, but, in the late 1980s, Hungarian party and government leaders joined public protests against this repression and the implementation of Romania's policy of reapportionment and relocation of the rural population, which affected a disproportionately large number of ethnic Hungarian settlements. A rift between Hungary and Romania deepened as ethnic disturbances in Transylvania continued even after the fall of the regime of Nicolae Ceauşescu in Romania. Hungary declared all Romanians of Hungarian descent to be dual citizens, in defiance of a 1979 bilateral agreement. This policy, combined with renewed openings toward Austria, establishment of trade relations with South Korea, and the resumption of diplomatic relations with Israel, was taken as a sign of a more independent foreign policy, as were the efforts at strengthening Hungary's ties with western Europe.

In 1989 the long-closed border between Hungary and Austria was opened. Interviews broadcast by Hungarian television with Alexander Dubček and Ota Šik, leaders of the Czechoslovak reform movement, and with the exiled king Michael of Romania, as well as Hungary's refusal to prevent tens of thousands of East German refugees from crossing into Austria on their way to West Germany, led to formal protests in Prague, Bucharest, and East Berlin.

The changes on the domestic scene were no less dramatic. Guidelines for a new constitution did not mention the "leading role of the Party," spelled out by the constitution of 1949. The draft of the new constitution sanctioned a multiparty system that had already been accepted in principle by the party leadership. The new constitution was based on the separation of legislative, executive, and judicial powers and included guarantees of individual and civil rights. Many of these changes were put into effect in October 1989 by a series of amendments to the 1949 constitution; other changes included the creation of the post of state president in place of the Presidential Council and the elimination of the word *People's* from the name of the country.

Constitutional amendments of 1989

Important new legislation included amendments to the law of assembly, which granted the holding of indoor meetings without special permission, and a new enterprise law, which allowed the private ownership of businesses with up to 500 employees, permitted foreigners to own up to 100 percent of an enterprise, and allowed mixed (*i.e.,* joint state and private) ownership of property. The government consulted with independent organizations and spokesmen of the opposition in the course of preparing the new laws, and the function of the National Assembly underwent a certain change. Theretofore a rubber stamp, the Assembly rejected the government's budget during its autumn 1988 session.

Alternative independent parties and organizations continued to grow. The Independent Smallholders' Party, the Social Democrats, the People's Party, and the Hungarian Independence Party, destroyed or emasculated by the communists in 1949, reemerged. A Christian Democratic Party also began to organize. Few of the new parties chose to support socialist principles. The same held true for the Young Hungarian Workers' Organization, the Federation of Agrarian Youth, the Alliance of Free Democrats (SzDSz), and the Hungarian Democratic Forum (MDF), a right-of-centre party whose membership passed 10,000 by the end of 1988. (G.Ba./Ed.)

Stabilizing democracy and the market economy. In March 1990 Hungary's first free elections led to a landslide victory for the opposition. The MDF, which had become the single strongest party in the parliament, formed a coalition government with the Smallholder and Christian Democratic parties under the premiership of József Antall (until his death in 1993). The MSzP, burdened by the legacy of state socialism, gained less than 10 percent of the votes. The coalition guided Hungary's socioeconomic transformation and attempted to stabilize the economy while implementing privatization and other elements of a market economy. The populist right wing of the MDF was vocal about the "national issue"—the question of the Hungarian minorities in neighbouring countries—and attempted to put it at the centre of the government's platform. After a period in which the discussion of anti-Semitism and discrimination against Gypsies had come to the fore, anti-Semitic rhetoric also appeared, as did an attempt to rehabilitate the reputation of the Horthy regime.

The new political elite, having miscalculated the impact of this policy and propaganda, antagonized a great part of the population. A deep dissatisfaction was expressed in the local elections a few months later that resulted in a devastating defeat for the MDF and a victory for the liberal-left opposition of the SzDSz and MSzP.

Economic transformation was more painful and lasted longer than anybody had thought it would. Unemployment jumped from nonexistence to 14 percent. Inflation

increased at an annual rate of 23–35 percent. The living standards of more than one-third of the populace declined to below subsistence level; income disparities increased; and corruption became more widespread and visible. Together with the previously omnipotent police force, street security collapsed in 1990, and the crime rate, especially in Budapest, increased threefold in five years.

In 1994, Hungary's second free elections transformed the country's politics. The MSzP won an absolute majority of the seats in the parliament and formed a government in coalition with the SzDSz under the premiership of Gyula Horn. Horn pursued many of the policies initiated by Antall, including the privatization of the economy and the pursuit of membership in NATO and the European Union (EU). In March 1995 the government introduced an austerity program designed to trim the budget and instill the financial discipline required by international financial institutions. The reforms precipitated several ministerial resignations, threatened the stability of the governing coalition, and were deeply unpopular.

The economy rebounded over the next several years, and the country attracted the largest amount of direct foreign investment in eastern and central Europe. Moreover, the Hungarian stock market became the region's strongest, resulting in a surge in the government's popularity.

In July 1997, Hungary, along with the Czech Republic and Poland, was invited to join NATO. The public overwhelmingly endorsed membership in November, and the country officially joined NATO in 1999. Meanwhile, the EU rated Hungary the best eastern and central European candidate for membership based on its commitment to democracy, its economic competitiveness, and its advancement in adjusting its laws to EU standards.

Hungarian politics slowly evolved into a bipolar system pitting the left against the centre-right. The centre-right Federation of Young Democrats (Fidesz)–Hungarian Civic Party, which promised improvements in the welfare system to ease the hardships caused by Horn's austerity program, won a surprising victory in May 1998. Fidesz formed a government by entering a coalition with the Hungarian Democratic Forum and the populist, right-wing Independent Smallholders' Party. Prime Minister Viktor Orban immediately launched a radical reform of state administration. He sought to enhance the power of the office of the prime minister, fired the boards that oversaw the social security funds and central social security payments, created a government newspaper, and purged many figures in the state-controlled media. As the country entered the 21st century, the economic boom continued, but the political situation deteriorated, particularly because of acrimonious relations between Orban and the opposition, which regarded the prime minister's leadership style as arrogant. In 2002, after a bitter election campaign, Orban's government was defeated by the Hungarian Socialist Party, led by economist Peter Medgyessy, who formed a two-party coalition. Medgyessy continued to pursue economic integration with western Europe and maintained cordial relations with the United States. In 2004 Hungary realized one of its primary foreign policy goals by gaining admission to the EU. (Ed.)

For later developments in the history of Hungary, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 921, 923, 961, 962, 963, and 972, and the *Index*.

BIBLIOGRAPHY

Physical and human geography. Overviews of the country are provided by STEPHEN R. BURANT (ed.), *Hungary: A Country Study*, 2nd ed. (1990); and STEPHEN BORSODY (ed.), *The Hungarians: A Divided Nation* (1988). MÁRTON PÉCSI and BÉLA SÁRFALVI, *The Geography of Hungary* (1964; originally published in Hungarian, 1960), gives a detailed picture of the physical and economic geography. MÁRTON PÉCSI, *Geomorphological Regions of Hungary*, trans. from Hungarian (1970), is a brief description of individual regions. Economic studies include IVAN T. BEREND (T. IVÁN BEREND) and GYÖRGY RÁNKI, *The Hungarian Economy in the Twentieth Century* (1985); and TIVADAR BERNÁT (ed.), *An Economic Geography of Hungary*, 2nd enlarged ed. (1989; originally published in Hungarian, 1969). LÓRÁNT CZIGÁNY, *The Oxford History of Hungarian Literature from the Earliest Times to the Present* (1984), provides critical studies. Hungary's cinematic achievements are detailed in GRAHAM PETRIE, *History Must Answer to Man: The Contemporary Hungarian Cinema* (1978).

History. Major histories in Hungarian include SÁNDOR SZILÁGYI (ed.), *A Magyar nemzet története*, 10 vol. (1895–98); BÁLINT HÓMAN and GYULA SZEKFŰ, *Magyar történet*, 7th ed., 5 vol. (1941–43); and ZSIGMOND PÁL PACH (ed.), *Magyarország története öt kötetben* (1976–). Major sections dealing with Hungary appear in ADAM WANDRUSZKA and PETER URBANITSCH (eds.), *Die Habsburgermonarchie, 1848–1918* (1973–). LADISLAS MAKKAI (LÁSZLÓ MAKKAI), *Histoire de Transylvanie* (1946), is also useful. General histories in English include DOMOKOS G. KOSÁRY, *A History of Hungary* (1941, reissued 1971); C.A. MACARTNEY, *Hungary, a Short History* (1962); PAUL IGNOTUS, *Hungary* (1972); IVAN T. BEREND (T. IVÁN BEREND) and GYÖRGY RÁNKY, *Hungary: A Century of Economic Development* (1974); ROBERT A. KANN, *A History of the Habsburg Empire, 1526–1918* (1974); PETER F. SUGAR, PÉTER HANÁK, and TIBOR FRANK (eds.), *A History of Hungary* (1990); and LÁSZLÓ KONTLER, *A History of Hungary: Millennium in Central Europe* (2002).

Early and medieval history is covered in C.A. MACARTNEY, *The Magyars in the Ninth Century* (1930, reprinted 1968); JENŐ SZŰCS, "The Three Historical Regions of Europe: An Outline," *Acta Historica Academiae Scientiarum Hungaricae*, 29(2–4):131–184 (1983); ZSIGMOND PÁL PACH, *Hungary and the European Economy in Early Modern Times* (1994); and JÁNOS M. BAK and BÉLA K. KIRÁLY (eds.), *From Hunyadi to Rákóczi: War and Society in Late Medieval and Early Modern Hungary* (1982).

The 18th and 19th centuries are discussed in: HENRY MARCZALI, *Hungary in the Eighteenth Century* (1910, reprinted 1971; originally published in Hungarian, 1882); DOMOKOS G. KOSÁRY, *Culture and Society in Eighteenth Century Hungary* (1987); C.A. MACARTNEY, *The Hapsburg Empire, 1790–1918* (1968); PÉTER HANÁK, *Ungarn in der Donaumonarchie* (1984); GEORGE BARANY, *Stephen Széchenyi and the Awakening of Hungarian Nationalism, 1791–1841* (1968); ISTVÁN DEÁK, *The Lawful Revolution: Louis Kossuth and the Hungarians, 1848–1849* (1979); GÁBOR VERMES, *István Tisza: The Liberal Vision and Conservative Statecraft of a Magyar Nationalist* (1985); R.W. SETON-WATSON, *Racial Problems in Hungary* (1908, reprinted 1972); and JOHN LUKACS, *Budapest 1900* (1988).

Studies of World War I and the Treaty of Trianon include MICHAEL KÁROLYI, *Fighting the World: The Struggle for Peace*, trans. from Hungarian (1923); OSCAR JÁSZI (OSZCAR JÁSZI), *The Dissolution of the Habsburg Monarchy* (1929, reprinted 1961); RUDOLF L. TŐKÉS, *Béla Kun and the Hungarian Soviet Republic: The Origins and Role of the Communist Party of Hungary in the Revolutions of 1918–1919* (1967); TIBOR HAJDU, *The Hungarian Soviet Republic* (1979; originally published in Hungarian, 1969); and C.A. MACARTNEY, *Hungary and Her Successors: The Treaty of Trianon and Its Consequences, 1919–1937* (1937, reprinted 1968).

Events of the years 1920–45 are detailed in IGNÁC ROMSICS, *István Bethlen: A Great Conservative Statesman of Hungary, 1874–1946*, trans. from Hungarian (1995); C.A. MACARTNEY, *A History of Hungary, 1929–1945*, 2 vol. (1956–57; also published as *October Fifteenth: A History of Modern Hungary, 1929–1945*, 2nd ed., 2 vol., 1961); THOMAS SAKMYSTER, *Hungary's Admiral on Horseback: Miklós Horthy, 1918–1944* (1994); MARIO D. FENYO, *Hitler, Horthy, and Hungary: German-Hungarian Relations, 1941–1944* (1972); GYULA JUHÁSZ, *Hungarian Foreign Policy, 1919–1945* (1979); RANDOLPH L. BRAHAM, *The Politics of Genocide: The Holocaust in Hungary*, rev. and enlarged ed., 2 vol. (1994); and ANDREW C. JANOS, *The Politics of Backwardness in Hungary, 1825–1945* (1982).

Hungary under communism is described in ERNST C. HELMREICH (ed.), *Hungary* (1957, reissued 1973); MIKLÓS MOLNÁR, *A Short History of the Hungarian Communist Party* (1978); BÉLA SZÁSZ, *Volunteers for the Gallows: Anatomy of a Show Trial*, trans. from Hungarian (1971); and BENNETT KOVRIG, *Communism in Hungary: From Kun to Kádár* (1979).

The period from the revolution of 1956 to the present is addressed in PAUL E. ZINNER, *Revolution in Hungary* (1962, reissued 1972); CHARLES GATI, *Hungary and the Soviet Bloc* (1986); IVAN T. BEREND (T. IVÁN BEREND), *The Hungarian Economic Reforms, 1953–1988* (1990; originally published in Hungarian, 1988), and *Central and Eastern Europe, 1944–1993* (1996), with major sections concerning Hungary; and PETER A. TOMA and IVAN VOLGYES, *Politics in Hungary* (1977). Further references may be found in STEVEN BELA VARDY, *Modern Hungarian Historiography* (1976); and GÉZA KILÉNYI and VANDA LAMM (eds.), *Democratic Changes in Hungary*, trans. from Hungarian (1990).

(C.A.M./G.Ba./I.T.B.)

The Hydrosphere

Water is the most abundant substance at the surface of the Earth. About 1.4 billion cubic kilometres (326 million cubic miles) of water in liquid and frozen form make up the oceans, lakes, streams, glaciers, and groundwaters found there. It is this enormous volume of water, in its various manifestations, that forms the discontinuous layer, enclosing much of the terrestrial surface, known as the hydrosphere.

Central to any discussion of the hydrosphere is the concept of the hydrologic cycle. This cycle consists of a group of reservoirs containing water, the processes by which water is transferred from one reservoir to another (or transformed from one state to another), and the rates of transfer associated with such processes. These transfer paths penetrate the entire hydrosphere, extending upward to about 15 kilometres (nine miles) in the Earth's atmosphere and downward to depths on the order of five kilometres in its crust.

This article examines the processes of the hydrologic cycle and discusses the way in which the various reservoirs of the hydrosphere are related through the hydrologic cycle. It also describes the biogeochemical properties of the waters of the Earth at some length and considers the distribution of global water resources and their utilization and pollution by human society. Details concerning the major water environments that make up the hydrosphere are provided in the articles OCEANS, LAKES, RIVERS, and ICE AND ICE FORMATIONS. See also CLIMATE AND WEATHER for specific information about the impact of climatic factors on the hydrologic cycle. The principal concerns and methods of hydrology and its various allied disciplines are summarized in EARTH SCIENCES, THE. For coverage of other related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 221, 222, 223, 351, and 10/33, and the *Index*.

The article is divided into the following sections:

Distribution and quantity of the Earth's waters 715
Biogeochemical properties of the hydrosphere 715
  Rainwater
  River and ocean waters
  Lake waters
  Groundwaters
  Ice
The hydrologic cycle 720
  General nature of the cycle
  Processes involved in the cycle
Origin and evolution of the hydrosphere 725
  The early hydrosphere
  The transitional hydrosphere
  The modern hydrosphere
Impact of human activities on the hydrosphere 728
  Eutrophication
  Acid rain
  Buildup of greenhouse gases
Bibliography 731

DISTRIBUTION AND QUANTITY OF THE EARTH'S WATERS

Ocean waters and waters trapped in the pore spaces of sediments make up most of the present-day hydrosphere (see Table 1). The total mass of water in the oceans equals about 50 percent of the mass of sedimentary rocks now in existence and about 5 percent of the mass of the Earth's crust as a whole. Deep and shallow groundwaters constitute a small percentage of the total water locked in the pores of sedimentary rocks—on the order of 3 to 15 percent. The amount of water in the atmosphere at any one time is trivial, equivalent to 0.013×10^6 cubic kilometres of liquid water, or about 0.001 percent of the total at the Earth's surface. This water, however, plays an important role in the water cycle.

At present, ice locks up a little more than 1 percent of the Earth's water and may have accounted for as much as 3 percent or more during the height of the glaciations of the Pleistocene epoch (from 1,600,000 to 10,000 years ago). Although water storage in rivers, lakes, and the atmosphere is small, the rate of water circulation through the rain–river–ocean–atmosphere system is relatively rapid. The amount of water discharged each year into the oceans from the land is approximately equal to the total mass of water stored at any instant in rivers and lakes.

Soil moisture accounts for only 0.005 percent of the water at the Earth's surface. It is this small amount of water, however, that exerts the most direct influence on evaporation from soils. The biosphere, though primarily H_2O in composition, contains very little of the total water at the terrestrial surface, only about 0.00004 percent. Yet, the biosphere plays a major role in the transport of water vapour back into the atmosphere by the process of transpiration.

As will be seen in the next section, the Earth's waters are not pure H_2O but contain dissolved and particulate materials. Thus, the masses of water at the Earth's surface are major receptacles of inorganic and organic substances, and water movement plays a dominant role in the transportation of these substances about the planet's surface.

Table 1: Water Masses at the Earth's Surface

reservoir	volume (in millions of cubic kilometres)	percent of total
Oceans	1,370	97.25
Ice caps and glaciers	29	2.05
Deep groundwater* (750–4,000 metres)	5.3	0.38
Shallow groundwater (< 750 metres)	4.2	0.30
Lakes	0.125	0.01
Soil moisture	0.065	0.005
Atmosphere†	0.013	0.001
Rivers	0.0017	0.0001
Biosphere	0.0006	0.00004
Total	1,408.7	100

*The total interstitial water in the pores of sediments is on the order of $50–300 \times 10^6$ km³. †As liquid equivalent of water vapour.
Source: Adapted from Elizabeth Kay Berner and Robert A. Berner, *The Global Water Cycle: Geochemistry and Environment* (1987), Prentice Hall, Inc.

BIOGEOCHEMICAL PROPERTIES OF THE HYDROSPHERE

Rainwater. About 110,300 cubic kilometres of rain fall on land each year. The total water in the atmosphere is 0.013×10^6 cubic kilometres, and this water, owing to precipitation and evaporation, turns over every 9.6 days. Rainwater is not pure but rather contains dissolved gases and salts, fine-ground particulate material, organic substances, and even bacteria. The sources of the materials in rainwater are the oceans, soils, fertilizers, air pollution, and fossil-fuel combustion.

Constituents of rainwater

Figure 1 shows some typical maps of the chloride and sodium content of rain over the continental United States. The concentration contours in Figure 1 are subparallel to the coastlines. It also has been observed that rains over oceanic islands and near coasts have ratios of major dissolved constituents very close to those found in seawa-

ter. The discovery of the high salt content of rain near coastlines was somewhat surprising because sea salts are not volatile, and it might be expected that the process of evaporation of water from the sea surface would "filter" out the salts. It has been demonstrated, however, that a large percentage of the salts in rain is derived from the bursting of small bubbles at the sea surface due to the impact of rain droplets or the breaking of waves, which results in the injection of sea aerosol into the atmosphere. This sea aerosol evaporates, with resultant precipitation of the salts as tiny particles that are subsequently carried high into the atmosphere by turbulent winds. These particles may then be transported over continents to fall in rain or as dry deposition.

Assuming equilibrium with the atmospheric carbon dioxide partial pressure (P_{CO_2}) of $10^{-3.5}$ atmosphere, the approximate mean composition of rainwater is in parts per million (ppm): sodium (Na^+), 1.98; potassium (K^+), 0.30; magnesium (Mg^{2+}), 0.27; calcium (Ca^{2+}), 0.09; chloride (Cl^-), 3.79; sulfate (SO_4^{2-}), 0.58; and bicarbonate (HCO_3^-), 0.12. In addition to these ions, rainwater contains small amounts of dissolved silica—about 0.30 ppm. The average pH value of rainwater is 5.7. (The term pH is defined as the negative logarithm of the hydrogen ion concentration in moles per litre. The pH scale ranges from 0 to 14, with lower numbers indicating increased acidity.) On a global basis, as much as 35 percent of the sodium, 55 percent of the chlorine, 15 percent of the potassium, and 37 percent of the sulfate in river water may be derived from the oceans through sea aerosol generation.

Significance of marine aerosols

A considerable amount of data has become available for marine aerosols. These aerosols are important because (1) they are vital to any description of the global biogeochemical cycle of an element; (2) they may have an impact on climate; (3) they are a sink, via heterogeneous chemical reactions, for trace atmospheric gases; and (4) they influence precipitation of cloud and rain droplets. For many trace metals the ratio of the atmospheric flux to the riverine flux for coastal and remote oceanic areas may be greater than one, indicating the importance of atmospheric transport. Figure 2 illustrates the enrichment factors (EF) of North Atlantic marine aerosols and suspended matter in North Atlantic waters relative to the crust, where

$$EF_{crust} = \frac{(X/Al)_{air}}{(X/Al)_{crust}},$$

and $(X/Al)_{air}$ and $(X/Al)_{crust}$ refer, respectively, to the ratio of the concentration of the element X to that of Al, aluminum, in the atmosphere and in average crustal material. The similarity in trend of enrichment factors for marine aerosols and suspended matter indicates qualitatively the importance of the marine aerosol to the composition of marine suspended matter and, consequently, to deep-sea sedimentation.

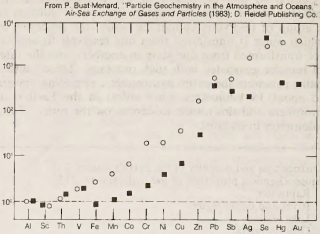


Figure 2: Element enrichment factors (EF) in North Atlantic marine aerosols (squares) compared with factors for suspended matter in North Atlantic ocean waters (circles).

Adapted from data from the National Atmospheric Deposition Program (NADP)

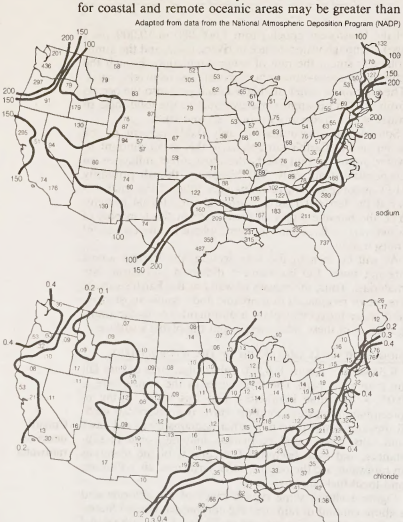


Figure 1: Contours of the average concentrations of dissolved sodium and chloride in precipitation over the United States during 1987.

In some instances the ratios of ions in rainwater deviate significantly from those in seawater. Mechanisms proposed for this fractionation are, for example, the escape of chlorine as gaseous hydrogen chloride (HCl) from sea-salt aerosol with a consequent enrichment in sodium and bubbling and thermal diffusion. In addition, release of gases like dimethyl sulfide (DMS) from the sea surface and its subsequent reaction in the oceanic atmosphere to sulfate can change rainwater ion ratios with respect to seawater. Soil particles also can influence rainwater composition. Rainfall over the southwestern United States contains relatively high sulfate concentrations because of sulfate-bearing particles that have been blown into the atmosphere from desert soils. Rain near industrial areas commonly contains high contents of sulfate, nitrate, and carbon dioxide (CO_2) largely derived from the burning of coal and oil. There are two main processes leading to the conversion of sulfur dioxide (SO_2) to sulfuric acid (H_2SO_4). These are reactions with hydroxyl radicals (OH) and with hydrogen peroxide (H_2O_2) in the atmosphere:

$$SO_2 + OH \rightarrow \text{intermediate species} \rightarrow H_2SO_4 \qquad (1)$$

and

$$SO_2 + H_2O_2 = H_2SO_4. \qquad (2)$$

The sulfuric acid then dissociates to hydrogen and sulfate ions:

$$H_2SO_4 = 2H^+ + SO_4^{2-}. \qquad (3)$$

For the nitrogen gases nitric oxide (NO) and nitrogen dioxide (NO_2) released from fossil-fuel burning, their atmospheric reactions lead to the production of nitric acid (HNO_3) and its dissociation to hydrogen ions (H^+) and nitrate (NO_3^-). These reactions are responsible for the acid rain conditions highly evident in the northeastern United States, southeastern Canada, and western Europe (see below *Acid rain*). The high sulfate values of the rain in the northeastern United States reflect the acid precipitation conditions of this region.

River and ocean waters. River discharge constitutes the main source for the oceans. Table 2 shows the difference between the composition of seawater and that of river, or

Difference in composition

stream, water. Seawater has a more uniform composition than river water. It contains, by weight, about 3.5 percent dissolved salts, whereas river water has only 0.012 percent. The average density of the world's oceans is roughly 2.75 percent greater than that of typical river water. Of the average 35 parts per thousand salts of seawater, sodium and chlorine make up almost 30 parts, and magnesium and sulfate contribute another four parts. Of the remaining one part of the salinity, calcium and potassium constitute 0.4 part each and carbon, as carbonate and bicarbonate, about 0.15 part. Thus, only eight elements (oxygen, sulfur, chlorine, sodium, magnesium, calcium, potassium, and carbon) make up 99 percent of seawater, though most of the 92 naturally occurring elements have been detected therein. Of importance are the nutrient elements phosphorus, nitrogen, and silicon, along with such essential micronutrient trace elements as iron, cobalt, and copper. These elements strongly regulate the organic production of the world's oceans.

In contrast to ocean water, the average salinity of the world's rivers is low—only about 0.012 percent, or 120 ppm by weight. Of this salt content, carbon as bicarbonate constitutes 58 parts, or 48 percent, and calcium, sulfur as sulfate, and silicon as dissolved monomeric silicic acid make up a total of about 39 parts, or 33 percent. The remaining 19 percent consists predominantly of chlorine, sodium, and magnesium in descending importance. It is obvious that the concentrations and relative proportions of dissolved species in river waters contrast sharply with those of seawater. Thus, even though seawater is derived in part by the chemical differentiation and evaporation of river water, the processes involved affect every element differently, indicating that simple evaporation and concentration are entirely secondary to other processes.

Generally speaking, the composition of river water, and thus that of lakes, is controlled by water–rock interactions. The attack of carbon dioxide-charged rain and soil waters on the individual minerals in continental rocks leads to the production of dissolved constituents for lakes, rivers, and streams. It also gives rise to solid alteration products that make up soils or suspended particles in freshwater aquatic systems. The carbon dioxide content of rain and soil waters is of particular importance in weathering processes. The pH of rainwater equilibrated with the atmospheric carbon dioxide partial pressure of $10^{-3.5}$ atmosphere is 5.7. In industrial regions, rainwater pH values may be lower because of the release and subsequent hydrolysis of acid gases—namely, sulfur dioxide and nitrogen oxides (NO_x) from the combustion of fossil fuels. After rainwater enters soils, its characteristics change markedly. The usual few parts per million of salts in rainwater increase substantially as the water reacts. The upper part of the soil is a zone of intense biochemical activity. The bacterial population near the surface is large, but it decreases rapidly downward with a steep gradient. One of the major biochemical processes of the bacteria is the oxidation of organic material, which

Table 2: Composition of Oceans and Streams

atomic number	atomic weight	element	seawater (in micrograms per litre or parts per billion)	streams* (in micrograms per litre or parts per billion)
1	1.00797	hydrogen	1.1×10^8	1.10×10^8
2	4.0026	helium	0.0068	†
3	6.939	lithium	180	3
4	9.0122	beryllium	0.0056	†
5	10.811	boron	4,440	18
6	12.01115	carbon (inorganic) and dissolved organic	28,000	11,620
7	14.0067	nitrogen (dissolved N_2, NO_3^-, NO_2^-, NO_4^+), and dissolved organic	500 15,000	10,800 950
8	15.9994	oxygen (dissolved O_2, H_2O)	6,000 8.83×10^8	† 8.83×10^8
9	18.9984	fluorine	1,300	100
10	20.183	neon	0.120	†
11	22.9898	sodium	1.08×10^7	7,200
12	24.312	magnesium	1.29×10^6	3,650
13	26.9815	aluminum	1	50
14	28.086	silicon	2,900	4,850
15	30.9738	phosphorus	60	78
16	32.064	sulfur	9.05×10^5	3,830
17	35.453	chlorine	1.9×10^7	8,250
18	39.948	argon	430	†
19	39.102	potassium	3.8×10^5	1,400
20	40.08	calcium	4.12×10^5	14,700
21	44.956	scandium	0.0006	0.004
22	47.90	titanium	1	10
23	50.942	vanadium	2.5	0.9
24	51.996	chromium	0.3	1
25	54.9380	manganese	0.2	8.2
26	55.847	iron	2	40
27	58.9332	cobalt	0.05	0.2
28	58.71	nickel	1.7	2.2
29	63.54	copper	0.5	10
30	65.37	zinc	4.9	30
31	69.72	gallium	0.03	0.09
32	72.59	germanium	0.05	†
33	74.9216	arsenic	3.7	2
34	78.96	selenium	0.2	0.2
35	79.909	bromine	67,300	20
36	83.80	krypton	0.21	†
37	85.47	rubidium	120	1.5
38	87.62	strontium	8,100	60
39	88.905	yttrium	0.013	0.07
40	91.22	zirconium	0.026	†
41	92.906	niobium	0.015	†
42	95.94	molybdenum	10	0.5
43	98.906	technetium	(not naturally occurring)	
44	101.07	ruthenium	0.0007	†
45	102.905	rhodium	†	†
46	106.4	palladium	†	†
47	107.87	silver	0.04	0.3
48	112.4	cadmium	0.11	†
49	114.82	indium	0.0001	†
50	118.69	tin	0.01	†
51	121.75	antimony	0.24	1
52	127.60	tellurium	†	†
53	126.90	iodine	64	7
54	131.30	xenon	0.05	†
55	132.91	cesium	0.40	0.035
56	137.34	barium	15	60
57	138.91	lanthanum	0.0034	0.05
58	140.12	cerium	0.0012	0.08
59	140.91	praseodymium	0.00064	0.007
60	144.24	neodymium	0.0028	0.04
61	(147)	promethium	(not naturally occurring)	
62	150.35	samarium	0.00045	0.008
63	151.96	europium	0.000130	0.001
64	157.25	gadolinium	0.00070	0.008
65	158.92	terbium	0.00014	0.001
66	162.50	dysprosium	0.00091	0.05
67	164.93	holmium	0.00022	0.001
68	167.26	erbium	0.00087	0.004
69	168.93	thulium	0.00017	0.001
70	173.04	ytterbium	0.00082	0.004
71	174.97	lutetium	0.00015	0.001
72	178.49	hafnium	<0.007	†
73	180.95	tantalum	<0.0025	†
74	183.85	tungsten	<0.1	0.03
75	186.2	rhenium	0.0084	†
76	190.2	osmium	†	†
77	192.2	iridium	†	†
78	195.09	platinum	†	†
79	196.97	gold	0.004	0.002
80	200.59	mercury	0.03	0.07
81	204.37	thallium	0.01	†
82	207.19	lead	0.03	1
83	208.98	bismuth	0.02	†
84–89 and 91		(thorium and uranium decay series elements: polonium, astatine, radon, francium, radium, actinium, and protactinium)		
90	232.04	thorium	<0.01	0.1
92	238.03	uranium	3.2	0.04

*Not corrected for anthropogenic influence. †Little data or no reasonable estimates available.
Source: Modified from R.M. Garrels, F.T. Mackenzie, and C. Hunt, *Chemical Cycles and the Global Environment*.
Copyright © 1975 by W. Kaufmann, Inc. All rights reserved.

leads to the release of carbon dioxide. Soil gases obtained above the zone of water saturation may contain 10 to 40 times as much carbon dioxide as the free atmosphere, and in some cases carbon dioxide has been shown to make up 30 percent of the soil gases as opposed to 0.03 percent of the free atmosphere. In addition to the acid effects of carbon dioxide, a highly acidic microenvironment is created by the roots of living plants. Values of pH as low as 2 have been measured immediately adjacent to root hairs. Plants may have several kilometres of root hairs, and so their chemical effects are formidable.

Congruent and incongruent weathering reactions

These acid solutions in the soil environment attack the rock minerals, the bases of the system, producing neutralization products of dissolved constituents and solid particles. Two general types of reaction occur: congruent and incongruent. In the former, a solid dissolves, adding elements to the water in their proportions in the mineral. An example of such a weathering reaction is the solution of calcite ($CaCO_3$) in limestones:

$$CaCO_3 + CO_2(g) + H_2O = Ca^{2+} + 2HCO_3^-. \quad (4)$$

Here, one of the HCO_3^- ions comes from calcite and the other from $CO_2(g)$ in the reacting water. The amount of carbon dioxide dissolved according to reaction (4) depends on temperature, pressure, original bicarbonate content of the weathering solution, and the partial pressure of the carbon dioxide. The carbon dioxide and the temperature are the most important variables. Increases in one or both of these variables lead to increases in the amount of calcite dissolved. For example, for a carbon dioxide pressure of $10^{-3.5}$ atmosphere, the amount of calcium that can be dissolved until saturation is about $10^{-3.3}$ mole, or 20 ppm, at 25° C (77° F). For an atmospheric carbon dioxide pressure of 10^{-2} atmosphere and for a soil atmosphere of nearly pure carbon dioxide, the values are 65 and 300 ppm, respectively. The weathering of calcite leads to the release of calcium and bicarbonate ions into soil waters and groundwaters, and these constituents eventually reach lake and river systems. The insoluble residue of quartz (SiO_2), clay minerals (cation, aluminum, silicon, oxygen, hydrogen phases), and iron oxides (*e.g.*, FeOOH) in the limestone rock make up the deep-red soils that form from limestone weathering. These particles may be carried into streams by runoff and hence to lakes and the oceans and become part of the suspended load of these systems.

An example of an incongruent weathering reaction is that involving aluminosilicates. One such reaction is the aggressive attack of carbon dioxide-charged soil water on the mineral K-spar ($KAlSi_3O_8$), an important phase found in continental rocks. The reaction is

$$2KAlSi_3O_8 + 2CO_2 + 11H_2O = Al_2Si_2O_5(OH)_4 + 2K^+ + 2HCO_3^- + 4H_4SiO_4. \quad (5)$$

It should be noted that the K-spar changes into a new mineral—kaolinite (a clay mineral) in this case—plus solution, and acid is consumed. The total dissolved material per litre of soil solution released is about 60 ppm for a solution initially equilibrated with a typical soil carbon dioxide content. The water resulting from reaction (5) would contain bicarbonate, potassium, and dissolved silica in the ratios 1:1:2, and the new solid, kaolinite, would be a weathering product. These dissolved constituents and the solid alteration product would eventually reach rivers to be transferred possibly to lakes and ultimately to the sea. It has been demonstrated that the composition of river water is the product of a variety of mineral–water reactions such as (4) and (5). The dissolved load of the world's rivers comes from the following sources: 7 percent from beds of halite (NaCl) and salt disseminated in rocks; 10 percent from gypsum ($CaSO_4 \cdot 2H_2O$) and anhydrite ($CaSO_4$) deposits and sulfate salts disseminated in rocks; 38 percent from limestones and dolomites; and 45 percent from the weathering of one silicate mineral to another. Of the bicarbonate ions in river water, 56 percent stems from the atmosphere, 35 percent from carbonate minerals, and 9 percent from the oxidative weathering of fossil organic matter. Reactions involving silicate minerals account for 30 percent of the riverine bicarbonate ions.

Besides dissolved substances, rivers also transport solids in traction (*i.e.*, bed load) and, most importantly, suspended load. The present global river-borne flux of solids to the oceans is estimated as 155×10^{14} grams per year. Most of this flux comes from Southeast Asian rivers. The composition of this suspended material resembles soils and shales and is dominated by silicon and aluminum. Present elemental fluxes are estimated in 10^{12} grams per year as silicon, 4,420; aluminum, 1,460; iron, 740; calcium, 330; potassium, 310; magnesium, 210; and sodium, 110. The total load of particulate organic carbon of the world's rivers is 180×10^{12} grams per year. The riverine fluxes of trace metals to the oceans are dominated by their occurrence in the particulate phase as opposed to the dissolved phase. The particulate matter in river water is an important source of silicon, aluminum, ion, titanium, rubidium, scandium, vanadium, the lanthanides, and other elements for deep-sea sediments.

Lake waters. Although lake waters constitute only a small percentage of the water in the hydrosphere, they are an important ephemeral storage reservoir for fresh water. Aside from their recreational use, lakes constitute a source of water for household, agricultural, and industrial uses. Lake waters are also very susceptible to changes in chemical composition due to these uses and to other factors.

In general, fresh waters at the continental surface evolve from their rock sources by enrichment in calcium and sodium and by depletion in magnesium and potassium. In very soft waters the alkalies may be more abundant than the alkaline earths, and in the more concentrated waters of open river systems $Ca > Mg > Na > K$. For the anions, in general, HCO_3^- exceeds SO_4^{2-}, which is greater in concentration than Cl^-. It is worthwhile at this stage to consider some major mechanisms that control global surface water composition. These mechanisms are atmospheric precipitation, rock reactions, and evaporation-precipitation. They are illustrated in Figure 3, which shows schematically the envelope of world surface water compo-

Mechanisms responsible for controlling surface water composition

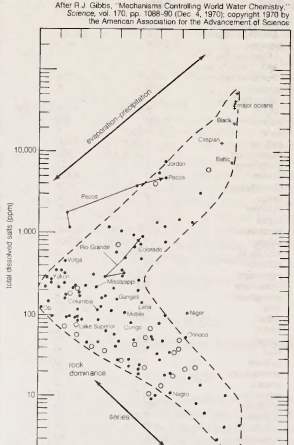


Figure 3: Variation in surface water composition as a function of total dissolved salt content and the ratio of dissolved sodium to sodium plus calcium in the waters. Three major mechanisms governing composition are shown in this diagram: evaporation-precipitation, precipitation, and rock–water reaction.

Table 3: Chemistry of Representative Closed-Basin, Saline Lake Waters*

	1 Carson Sink, Nevada	2 Bristol Lake, California	3 Salton Sea, California	4 Great Salt Lake, Utah	5 Surprise Valley Lake, California	6 Deep Springs Lake, California
SiO_2	19	—	20.8	48	36	—
Ca	261	43,296	505	241	11	3.1
Mg	129	1,061	581	7,200	31	1.2
Na	56,800	57,365	6,249	83,600	4,090	111,000
K	3,240	3,294	112	4,070	11	19,500
HCO_3	322	—	232	251	1,410	9,360
CO_3	—	—	—	—	664	22,000
SO_4	786	223	4,139	16,400	900	57,100
Cl	88,900	172,933	9,033	140,000	4,110	119,000
Total	152,000	279,150	20,900	254,000	10,600	335,000
pH	7.8			7.4	9.2	

*In parts per million.
Source: Modified from J.I. Drever, *The Geochemistry of Natural Waters*. © 1982. Prentice Hall, Inc.

sitions in terms of the variables of total dissolved salts and $Na/(Na + Ca)$ ratio.

The mechanism principally responsible for waters of very low salinity is precipitation. These waters tend to form in tropical regions of low relief and thoroughly leached source rocks. In these regions rainfall is high, and water compositions are usually dominated by salts brought in by precipitation. Such waters constitute one end-member of a series of water compositions for which the other end-member represents water compositions dominated by contributions of dissolved salts from the rocks and soils of their basins. These waters have moderate salinity and are rich in dissolved calcium and bicarbonate. They are, in turn, the end-member of another series that extends from the calcium-rich, medium-salinity fresh waters to the high-salinity, sodium chloride-dominated waters of which seawater is an example. Seawater composition, however, does not evolve directly from the composition of fresh waters and the precipitation of calcium carbonate; other mechanisms that control its composition are involved. Such factors as relief and vegetation also may affect the composition of the world's surface waters, but atmospheric precipitation, water–rock reactions, and evaporation–crystallization processes appear to be the dominant mechanisms governing continental surface water chemistry.

Continental fresh waters evaporate once they have entered closed basins, and their constituent salts precipitate on the basin floors. The composition of these waters may evolve along several different paths, depending on their initial chemical makeup. Table 3 shows a number of brine compositions from North American saline lakes, and Figure 4 illustrates possible evaporation paths that led to these compositions. For example, Surprise Valley Lake, California, is a body of sodium chloride water that may have evolved from what was initially fresh water which precipitated calcite during evaporation, resulting in water with a dissolved-calcium concentration to alkalinity ratio

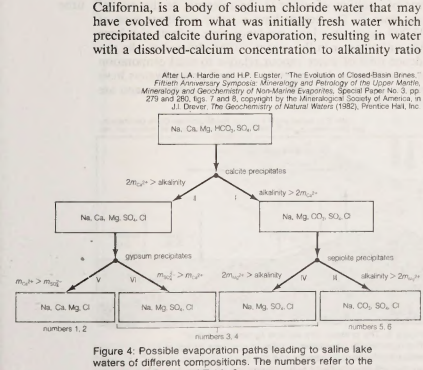


Figure 4: Possible evaporation paths leading to saline lake waters of different compositions. The numbers refer to the water compositions of Table 3.

of less than 2. Further evaporation of such water and precipitation of sepiolite $[Mg_2Si_3O_6(OH)_2]$ would give rise to water enriched in alkalinity with respect to dissolved magnesium and hence to sodium chloride–bicarbonate–carbonate water such as that of Surprise Valley Lake. The paths shown in Figure 4 are an oversimplification of the actual processes, which are far more diverse and complex in nature, but they do provide a reasonable picture of how fresh waters evolve into saline lake waters in closed basins.

Biological processes strongly affect the composition of lake waters and are responsible to a significant degree for the compositional differences between the upper water layer (the epilimnion) and the lower water layer (the hypolimnion) of lakes. The starting point is photosynthesis, represented by the following reaction:

$$106CO_2 + 16NO_3^- + HPO_4^{2-} + 122H_2O + 18H^+ \xrightarrow[\text{micro-nutrients}]{\text{radiant energy}} C_{106}H_{263}O_{110}N_{16}P + 138O_2. \quad (6)$$

The reversal of this reaction is oxidation-respiration leading to the release of the nutrients nitrogen and phosphorus, as well as carbon dioxide. In a stratified lake, carbon, nutrients, and silica are extracted from the epilimnion during photosynthesis. This process leads to reduced concentrations of nitrate, phosphate, and silica in these waters and, during times of maximum daylight organic production, supersaturation of epilimnion waters with respect to dissolved oxygen. The organic matter produced by phytoplankton may be either grazed upon by zooplankton and other organisms or decomposed by bacteria. Some of it, however, sinks into the hypolimnion. There, it is further decomposed, especially by bacteria, resulting in the release of dissolved phosphorus and nitrogen and the consumption of oxygen. Oxygen concentrations therefore are reduced in these lower lake waters, because stratification prevents oxygen exchange with the atmosphere. Furthermore, the inorganic carbonate and siliceous skeletons of the dead organisms sinking into the hypolimnion may dissolve, giving rise to increased concentrations of dissolved silica and inorganic carbon in the deep waters of stratified lakes. This dissolution is a result of undersaturation of the hypolimnion waters with respect to the opaline silica and calcium carbonate that make up the skeletons of the dead and sinking plankton. These natural biological processes have been accelerated in some lakes because of excess nutrient input by human activity, resulting in the eutrophication of lake waters (and marine systems; see below *Eutrophication*).

Groundwaters. These waters derive their compositions from a variety of processes, including dissolution, hydrolysis, and precipitation reactions; adsorption and ion exchange; oxidation and reduction; gas exchange between groundwater and the atmosphere; and biological processes (see Table 4). The biological processes of greatest importance are microbial metabolism, organic production, and respiration (oxidation). By far the most important overall process for the major constituents of groundwater is that of mineral–water reactions, which were briefly described

Processes affecting groundwater composition

above in *River and ocean waters*. Thus, the composition of groundwaters strongly reflects the types of rock minerals that the waters have encountered in their movement through the subsurface. Table 5 shows the waters found in limestones, crystalline rocks, and two types of fine-grained rocks. The mineralogy of the Wissahickon schist is dominated by aluminosilicate compositions, whereas the Ecca shale contains significant carbonate and dispersed salts. The latter minerals are more soluble than aluminosilicate minerals, and their dissolution gives rise to the high salinity of Ecca shale waters. The waters of the Wissahickon schist have low salinity partly because of the low chemical reactivity of the silicate minerals in this rock. In contrast, the Miocene limestone waters are dominated by dissolved calcium and bicarbonate, a characteristic reflecting the higher solubility and rate of dissolution of the calcite that makes up this rock. The groundwater from the granite is rich in the Ca^{2+} and Na^{+} cations derived from dissolution of the plagioclase feldspar (a sodium, calcium aluminosilicate), which is a major mineral found in this rock type. In general, the most mobile elements in groundwater—*i.e.*, those most easily liberated by the weathering of rock minerals—are calcium, sodium, and magnesium. Silicon and potassium have intermediate mobilities, and aluminum and iron are essentially immobile and locked up in solid phases.

Table 4: Processes Affecting the Major Chemical Components of Groundwater

component	origin*
Na^{+}	sodium chloride dissolution (some pollutive) plagioclase weathering rainwater addition
K^{+}	biotite weathering K-feldspar weathering biomass decreases dissolution of trapped aerosols
Mg^{2+}	amphibole and pyroxene weathering biotite (and chlorite) weathering dolomite weathering olivine weathering rainwater addition
Ca^{2+}	calcite weathering plagioclase weathering dolomite weathering dissolution of trapped aerosols biomass decreases
HCO_3^{-}	calcite and dolomite weathering silicate weathering
SO_4^{2-}	pyrite weathering (some pollutive) calcium sulfate dissolution rainwater addition
Cl^{-}	sodium chloride dissolution (some pollutive) rainwater addition
H_4SiO_4	silicate weathering

*The sources for each constituent are given in approximate order of decreasing importance.
Source: Adapted from Elizabeth Kay Berner and Robert A. Berner, *The Global Water Cycle: Geochemistry and Environment*. © 1987. Prentice Hall, Inc.

Groundwaters are highly susceptible to contamination because of human activities and the fact that their dissolved constituents are derived to a large extent from the leaching of surface materials. Some of the nitrogen and phosphorus applied to soils as fertilizers and organic pesticides may be leached and leak into groundwater systems, leading to increased concentrations of ammonium and phosphate. Radioactive wastes, industrial chemicals, household materials, and mine refuse are other anthropogenic sources of dissolved substances that have been detected in groundwater systems.

Ice. Ice is nearly a pure solid and, as such, accommodates few foreign ions in its structure. It does contain, however, particulate matter and gases, which are trapped in bubbles within the ice. The change in composition of these materials through time, as recorded in the successive layers of ice, has been used to interpret the history of the Earth's surface environment and the impact of human activities on this environment. The increase in the lead content of continental glacial ice with decreasing age of the ice, for example, reflects the progressive input of tetraethyl lead into the global environment from gasoline burning. Also, atmospheric carbon dioxide and methane concentrations, which have increased significantly during the past century because of anthropogenic activities, are faithfully recorded in ice bubbles of the thick continental ice sheets. Present-day atmospheric carbon dioxide and methane concentrations are 25 percent and 167 percent, respectively, higher than their concentrations 200 years ago; the latter concentration values were obtained from measurements of the gases in air trapped in ice.

Table 5: Groundwater Compositions in Major Rock Types

	limy	argillaceous	salty	crystalline
	Miocene limestone	Wissahickon schist	Ecca shale, South Carolina	granite, South Carolina
SiO_2	8.9	14.0	32	35.0
Ca^{2+}	48.0	3.1	62	13.0
Mg^{2+}	5.8	1.2	64	4.3
Na^{+}	4.0	3.3	92	8.4
K^{+}	0.7	0.8		3.5
HCO_3^{-}	168.0	21.0	362	72.0
SO_4^{2-}	6.4	1.2	106	6.9
Cl^{-}	4.8	2.4	140	3.8
Total	246.6	47.0	858	146.9

Source: R.M. Garrels and F.T. Mackenzie, *Evolution of Sedimentary Rocks* (1971). W.W. Norton and Co.

THE HYDROLOGIC CYCLE

General nature of the cycle. The present-day hydrologic cycle at the Earth's surface is illustrated in Figure 5. Some 496 cubic kilometres of water evaporate from the land and ocean surface annually, remaining for about 10 days in the atmosphere before falling as rain or snow. The amount of solar radiation necessary to evaporate this water is half of the total solar radiation received at the Earth's surface. About one-third of the precipitation falling on land runs off to the oceans primarily in rivers, while direct groundwater discharge to the oceans accounts for only about 6 percent of the total discharge. A small amount of precipitation is temporarily stored in the waters of rivers and lakes. The remaining precipitation over land, 0.073×10^6 cubic kilometres per year, returns to the atmosphere by evaporation. Over the oceans, evaporation exceeds precipitation, and the net difference represents transport of water vapour over land, where it precipitates as rain and returns to the oceans as river runoff and direct groundwater discharge.

Water residence time

The various reservoirs in the hydrologic cycle have different water residence times. Residence time is defined as the amount of water in a reservoir divided by either the rate of addition of water to the reservoir or the rate of loss from it. The oceans have a water residence time of 37,000 years; this long residence time reflects the large amount of water in the oceans. In the atmosphere the residence time of water vapour relative to total evaporation is only 10 days. Lakes, rivers, ice, and groundwaters have residence times lying between these two extremes and are highly variable.

After E.K. Berner and R.A. Berner, *The Global Water Cycle: Geochemistry and Environment*, p. 14. © 1987, reprinted by permission of Prentice Hall, Inc.

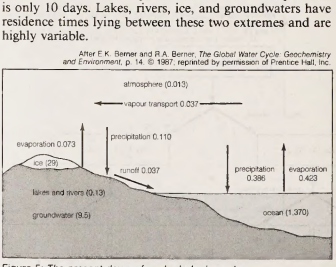


Figure 5: *The present-day surface hydrologic cycle.* The numbers in parentheses refer to volumes of water in millions of cubic kilometres, and the fluxes adjacent to the arrows are in millions of cubic kilometres of water per year.

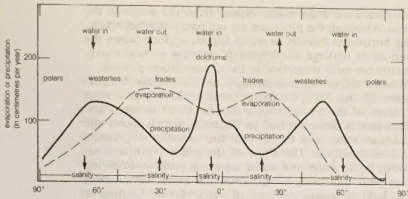


Figure 6: Latitudinal variation in precipitation and evaporation and relationship to major wind belts and oceanic salinity.
After R.M. Garrels, F.T. Mackenzie, and C. Hunt, *Chemical Cycles and the Global Environment* (1975); W. Kaufmann, Inc.

There is considerable variation in evaporation and precipitation over the globe. In order to have precipitation, there must be sufficient atmospheric water vapour and enough rising air to carry the vapour to an altitude where it can condense and precipitate. Figure 6 shows the latitudinal variation of precipitation and evaporation and their gross relation to the global wind belts. The trade winds, for example, are initially cool, but they warm up as they blow toward the equator. These winds pick up moisture from the ocean, increasing ocean surface salinity and causing seawater to sink. When the trade winds reach the equator, they rise, and the water vapour in them condenses and forms clouds. Net precipitation is high near the equator and also in the belts of the prevailing westerlies where there is frequent storm activity. Evaporation exceeds precipitation in the subtropics where the air is stable and near the poles where the air is both stable and has a low water vapour content because of the cold. The Greenland and Antarctic ice sheets formed due to the very low evaporation rates at the poles that result in precipitation exceeding evaporation in these local regions. The strong link between wind belts, the water balance, and the salinity of ocean water is apparent in Figure 6.

Processes involved in the cycle. The hydrologic cycle consists of various complicated processes of precipitation, evaporation, interception, transpiration, infiltration, percolation, retention, detention, overland flow, throughflow, and runoff. Figure 7 provides a schematic representation of the details of the hydrologic cycle, illustrating the processes involved. In the following sections, some of these processes are discussed in detail.

Water vapour and precipitation. As noted above, water exists in the atmosphere in gaseous form. Its liquid form, either as water droplets in clouds or as rain, and its solid form, as ice crystals in clouds, snowflakes, or hail, occur only momentarily and locally.

Primary functions of water vapour

Water vapour performs two major functions: (1) it is important to the radiation balance of the Earth as its presence keeps the planetary surface warmer than would otherwise be the case; and (2) it is the principal phase of the ascending part of the hydrologic cycle.

The mass of water vapour in the atmosphere, which represents only 0.001 percent of the hydrosphere, is highest in the tropics and decreases toward the poles. At a mean temperature of the Earth's surface of 15° C, the partial pressure of water vapour at equilibrium with pure water is 0.017 atmosphere. The addition of salts to pure water lowers its vapour pressure. The equilibrium, or saturation, water vapour pressure of a saturated solution of sodium chloride is 22 percent lower than that of pure water. Precipitable water vapour has, on the average, a vapour pressure of 0.0025 atmosphere, which amounts to 15 percent of the saturation vapour pressure. The ratio of observed water vapour pressure to the saturation vapour pressure at the same temperature is the relative humidity of the air. Thus, the mean relative humidity of the atmosphere is only 15 percent, a value that is low in human and biological terms. The relative humidity of the air, however, varies greatly from one geographic region to another and also vertically in the atmosphere. Atmospheric water vapour decreases rapidly with increasing altitude relative to its surface value. The amount of water required to saturate a volume of air depends on the temperature of the air. Air at high temperature can hold more water vapour at saturation than can air at low temperature. Because the temperature of the lower atmosphere (the troposphere) decreases rapidly with increasing altitude to about 15 kilometres, the upper levels of the troposphere contain little water vapour; most of the vapour is found within a few of kilometres of the Earth's surface. The average relative humidity of tropospheric air is about 50 percent. Above 15 kilometres, water vapour is essentially frozen out of

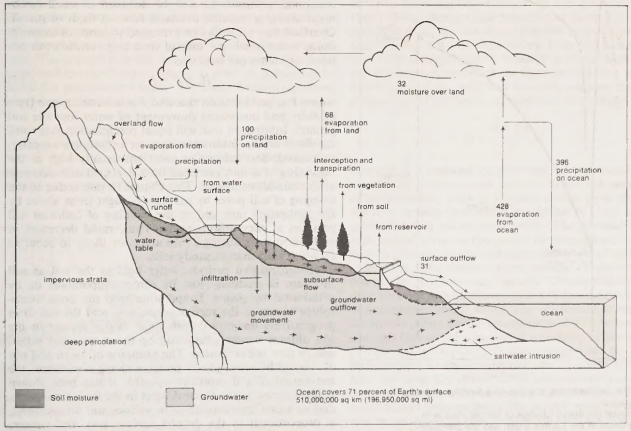


Figure 7: The detailed hydrologic cycle emphasizing processes involved in the transfer of water in the cycle. Numbers on arrows show relative water fluxes.

the atmosphere, amounting to less than 0.1 percent of its concentration at the Earth's surface.

Aside from temperature, other factors determine the water vapour content of the air and are particularly important in the lower troposphere. These factors include local evaporation and the horizontal atmospheric transportation of moisture, which varies with altitude, latitude, season, and topography. During a period of 10 days (*i.e.*, the residence time of water in the atmosphere), horizontal eddy turbulence may disperse vapour over distances up to 1,000 kilometres.

Interaction between air masses and surface water bodies

When a mass of air at the Earth's surface is exposed to a body of water, it gains water by evaporation or loses water by precipitation, depending on its relative humidity. If the air is undersaturated, with a relative humidity of less than 100 percent, it gains water vapour because the rate of evaporation exceeds the rate of condensation. If the air is supersaturated, with a relative humidity greater than 100 percent, the air mass loses water vapour because the rate of precipitation exceeds that of evaporation. This interaction between air masses and surface water bodies drives the atmosphere toward a state of saturation, which is not achieved for the entire atmosphere because of the variability in weather and because not all air masses are in contact with water bodies. In general, the level of atmospheric water vapour is higher in the summer, since temperatures are higher at this time of year. Also, atmospheric water vapour content is higher near the source of moisture than in distant regions. Over the oceans, the air is almost always near saturation, whereas over the deserts, where the supply of moisture is limited, the air is far below water vapour saturation values. In most cases, atmospheric water vapour content decreases inland over continents, but this decrease is modified by rainfall conditions, by the presence or absence of high mountains, large lakes, extensive forests, and swamps, and by the prevailing wind directions. Horizontal winds and air mass movements transfer water vapour from the ocean to the land. Although the processes are not completely separable, the horizontal transfer of water vapour seldom causes the vapour to undergo condensation, whereas vertical movements are most important in the condensation process.

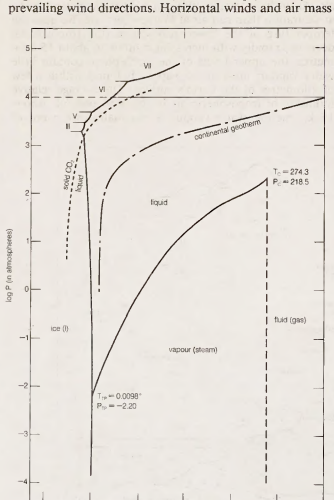


Figure 8: *The various states of water as a function of pressure and temperature.*
The continental geotherm (change of temperature with increasing depth in the Earth's crust) and the pressure–temperature boundary for solid CO_2–liquid CO_2 are shown for comparison.

Condensation depends strongly on the average temperature of the Earth's surface because the water vapour content of the air is strongly dependent on temperature. This fact is demonstrated in Figure 8, which shows the states of water as a function of the variables of pressure and temperature. The slope of the phase boundary between liquid water and water vapour is positive, implying that with increasing temperature the air at equilibrium will hold increasing amounts of water vapour. Cooling or mixing of this air results in condensation of the vapour and precipitation as water droplets or as ice crystals if the air temperature is below 0° C. When first formed, the water droplets or ice crystals are very small, on the order of 10^{-2} to 10^{-3} centimetre in diameter, and they float freely in the atmosphere. In large quantities, these water droplets and ice crystals produce a cloud. All clouds are formed as a result of cooling below the dew point, the temperature at which condensation begins when air is cooled at constant pressure and constant water vapour content. When the droplets or crystals coalesce to a size of about 10^{-2} centimetre in diameter, the drops become heavy enough to fall as raindrops or snowflakes. Hailstones measure about 10^{-1} centimetre in diameter or much larger. Water vapour condensing in the atmosphere contains strongly soluble salts (mostly of oceanic origin), weakly soluble or insoluble solids (dust), and dissolved gases. The dust and sea-salt aerosol particles in the air may act as sites of condensation by serving as nuclei for bringing initially a few water molecules together and inducing condensation from supersaturated air.

Distribution of precipitation. Precipitation falling toward the Earth's surface may suffer several fates. It may be evaporated during its fall or after it reaches the ground surface. If the surface is covered with dense vegetation, much of the precipitation may be held on leaves and plant limbs and stems. This process is termed interception and may result in little water reaching the ground because the water may be directly evaporated from plant surfaces back into the atmosphere. If precipitation reaches the ground in the form of snow, it may remain there for some time. On the other hand, if precipitation falls as rain, it may evaporate, infiltrate the soil, be detained in small catchment areas, or become overland flow—a form of runoff. Overland flow (R_o) may be expressed in terms of intensity units, water depth per unit of time (*e.g.*, centimetres per hour, or inches per hour), as

Overland flow

$$R_o = P - I, \tag{7}$$

where P is precipitation rate and I is infiltration rate (rate of entry and downward movement of water into the soil profile). Infiltration rate will equal precipitation rate until the limit of the infiltration rate, or infiltration capacity, is reached. Soil infiltration rates are usually high at the beginning of a rain preceded by a dry spell and decrease as the rainfall continues. This change in rate is due to the clogging of soil pores by particles brought from above by the infiltrating rain and to the swelling of colloidal soil particles as they absorb water. Thus, rapid decreases in infiltration rates during a rain are more likely to occur in clay-rich soils than in sandy soils.

Between rainfall periods, water held in the soil as soil moisture is gradually lost by direct evaporation or by withdrawal by plants. Evaporation into the open atmosphere occurs at the surface of the soil, and the soil dries progressively downward with time. Water vapour in the soil diffuses upward, replenishing the evaporated water, and in turn is evaporated. The pumping of air in and out of the soil by atmospheric pressure changes enhances the movement of soil moisture upward. It has been shown that evaporation of a water droplet in the free atmosphere, and to a first approximation in various soil atmospheres, is proportional to the droplet surface area $4\pi r^2$ (square centimetres), the diffusional flux of water at the droplet surface, and the transfer of heat as the droplet evaporates.

The equation for the rate of shrinkage of a water droplet due to evaporation is

$$\frac{dr}{dt} = -\frac{Dp_{v_0}(Sp-1)}{rp_L(1+X)}, \tag{8}$$

where dr/dt is the rate of change in the radius of the water droplet (centimetres per second), D is the diffusion coefficient of water vapour in air (cubic centimetres per second), ρ_{v_0} is the equilibrium vapour concentration at the droplet surface, Sp is the degree of undersaturation of water vapour in the environment, r is the radius of the droplet (centimetres), ρ_L is the density of liquid water (grams per cubic centimetre), and X is a dimensionless parameter depending on D, ρ_{v_0}, temperature, the heat of evaporation of water vapour, the coefficient of thermal conductivity of air, and the spherical coordinate system necessary to define processes occurring to a spherical water droplet. Water droplets shrink—$dr/dt < 0$, evaporate—when the water vapour concentration in the environment (atmosphere or soil atmosphere) is less than the saturation water vapour concentration at the droplet surface. They grow—$dr/dt > 0$, condense—when the converse is true in the free atmosphere. The term dr/dt has negative values for evaporation and positive ones for condensation. Use of this equation shows, as an example, that it would take 23 minutes for a water droplet to shrink (evaporate) in size from 50 to five micrometres in air at 10° C and a water vapour undersaturation of 1 percent.

Besides simple evaporation of water from soils, water is also returned to the atmosphere by transpiration in plants. Plants draw water from soil moisture through their vast network of root hairs and rootlets. This water is carried upward through the plant trunk and branches into the leaves, where it is discharged as water vapour.

Evapotranspiration

The term evapotranspiration is used in climatic and hydrologic studies to include the combined water loss from the Earth's surface resulting from evaporation and transpiration. The maximum possible evapotranspiration is termed potential evapotranspiration and is governed by the available heat energy. It is taken as equal to evaporation from a large water surface and is generally much less than actual evapotranspiration. Actual evapotranspiration is never greater than precipitation except on irrigated land because of percolation of water into groundwater bodies and surface runoff.

Figure 9 provides a detailed picture of the hydrologic cycle in the soil zone and at depth. The soil-moisture zone gains water by precipitation and infiltration and loses water by evapotranspiration, overland flow, and percolation of water downward due to gravity into the groundwater zone. The contact between the groundwater zone (phreatic zone) and the overlying unsaturated zone (vadose zone) is called the groundwater table. The water balance equation for change of soil-moisture storage in a soil is given as

$$S = P - E - R, \tag{9}$$

where S is storage, P is precipitation, E is evaporation, and R is surface runoff plus percolation rate into the groundwater zone; all terms are in units of length per unit of time (*e.g.*, millimetres per day, centimetres per month).

After A.N. Strahler and A.H. Strahler, *Environmental Geoscience* (1973); Hamilton Publishing Co.

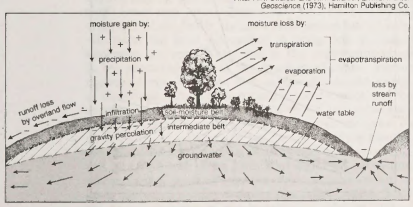


Figure 9: *The cycle of water in the soil zone and at depth.* Many factors influence water movement in this surface zone and are illustrated here.

In humid, midlatitude climates where a strong contrast between winter and summer temperatures exists, there is an annual cycle of the water content of soils. One such annual cycle is shown in Figure 10 and demonstrates the processes controlling soil moisture. Of special importance is the fact that the soils are saturated in this temperate climate in the spring, and the evaporation rate is low because of the low input of radiant energy from the Sun. By contrast, in the summer, evaporation increases because of increasing solar radiation, and with the growth of plants so does transpiration. Soil moisture is reduced to very low levels at this time of year.

After A.N. Strahler and A.H. Strahler, *Environmental Geoscience* (1973); Hamilton Publishing Co.

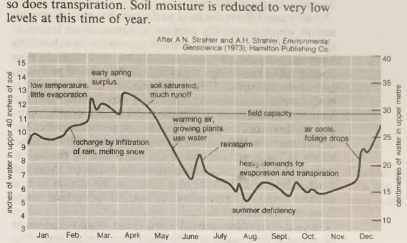


Figure 10: Annual cycle of moisture in soil in Ohio, U.S., showing factors controlling soil moisture content.

Groundwaters and river runoff. This section focuses on term R in equation (9) representing groundwater and river runoff losses from the soil-moisture zone. Water percolates from the soil-moisture zone through the unsaturated (vadose) zone to the water table. Flow through the unsaturated zone is complicated. After a rainfall, water may form nearly a continuous phase in pores in this zone, but with drying the last amount of water is held in wedges at points of contact of solid grains and as thin films on solid surfaces. The flow paths of water become more tortuous, and the water-conducting properties decrease rapidly. Structured soils and fractured rock in the vadose zone may act as conduits for fluids to reach the water table. Because of the complex geometry of water contained in the unsaturated zone, the properties of water are expressed by means of empirical relationships. Darcy's law, derived in 1856 from experimentation by the French engineer Henri Darcy, permits quantification of water flow through porous media. The law states that the rate of flow Q of a fluid through a porous layer of medium (*e.g.*, a sand bed) is directly proportional to the area A of the layer and to the difference Δh between the fluid heads at the inlet and outlet faces of the layer and inversely proportional to the thickness L of the layer, or, expressed analytically,

Darcy's law

$$Q = \frac{KA\,\Delta h}{L}, \tag{10}$$

where K is a constant characteristic of the medium. The term K for a porous rock medium is the volume of fluid of unit viscosity passing through a unit cross section of the rock in unit time under the action of a unit pressure gradient and is called permeability. The permeability of a rock is dependent on the geometric properties of the rock, such as porosity, shape and size distribution of constituent rock grains, and degree of cementation of the rock. Permeabilities of rocks vary greatly. Unconsolidated sands may have permeabilities measured in hundreds of darcys, whereas consolidated sands that will transmit reasonable amounts of fluid have permeabilities of 0.01 to one darcy. A rough idea of the meaning of one darcy of permeability (which equals 9.869×10^{-13} square metre) can be obtained by imagining a cube of sand one foot on a side. If the sand has a permeability of one darcy, approximately one barrel of water per day will pass through the one-foot cube with a one-pound pressure head. The general equation of Darcy can be modified to express flow in both the unsaturated zone and the saturated groundwater zone.

Groundwater is constantly in motion. When a lake

or stream intersects the groundwater table, groundwater communicates directly with these bodies of water. If the groundwater table is higher than the stream or lake level, a pressure head will develop such that the groundwater flows into the water body; conversely, if the groundwater table is lower than the river or lake level, the pressure gradient induces flow into the groundwater. Most groundwater ultimately reaches the channels of surface streams and rivers and flows to the sea. On the average, groundwater contributes to total river runoff about 30 percent of its water on a global basis.

Water runoff from the land surface is that part of precipitation which eventually appears in perennial or intermittent surface streams. Streamflow-generation mechanisms have been studied for several decades, and there is now considerable knowledge regarding rainfall-runoff processes and their controls. This understanding is the result of both careful observations from field experiments and the heuristic simulations of hypothetical realities with rigorous mathematical models. The discharge measured at the downstream end of a channel reach is supplied by channel inflow at the upstream end of the reach and by the lateral inflows that enter the channel from the hillslope along the reach. The lateral inflows may arrive at the stream in one of three forms: (1) groundwater flow, (2) subsurface storm flow, or (3) overland flow.

Groundwater flow provides the base flow component of streams that sustains their flow between storms. The "flashy" response of streamflow to individual precipitation events may be ascribed to either subsurface storm flow or overland flow. Subsurface storm flow can be a dominant streamflow-generation mechanism only when the impeding subsoil horizon laterally diverts infiltrating water downslope. Under intense rainfall events during which the surface soil layer becomes saturated to some depth, water is able to migrate through "preferred pathways" rapidly enough to deliver contributions to the stream during the peak runoff period. The conditions for subsurface storm flow are quite restrictive. The mechanism is most likely to be operative on steep, humid, forested hillslopes with very permeable surface soils.

Overland flow is generated at a point on a hillslope only after surface ponding takes place. Ponding cannot occur until the surface soil layers become saturated. It is now widely recognized that surface saturation can occur because of two quite distinct mechanisms—namely, Horton overland flow and Dunne overland flow.

Horton overland flow

The former classic mechanism is for a precipitation rate that exceeds the saturated hydraulic conductivity of the surface soil. A moisture content versus depth profile during such a rainfall event will show moisture contents that increase at the surface as a function of time. At some point in time the surface becomes saturated, and an inverted zone of saturation begins to propagate downward into the soil. It is at this time that the infiltration rate drops below the rainfall rate and overland flow is generated. The time is called the ponding time. The necessary conditions for the generation of overland flow by the Horton mechanism are (1) a rainfall rate greater than the saturated hydraulic conductivity of the soil and (2) a rainfall duration longer than the required ponding time for a given initial moisture profile. Horton overland flow is generated from partial areas of the hillslope where surface hydraulic conductivities are lowest.

Dunne overland flow

In Dunne overland flow, the precipitation rate is less than the saturated hydraulic conductivity, and the initial water table is shallow or there is a shallow impeding layer. Surface saturation occurs because of a rising water table; ponding and overland flow occur at a time when no further soil-moisture storage is available. The Dunne mechanism is more common to near-channel areas. Dunne overland flow is generated from partial areas of the hillslope where water tables are shallowest. Both Horton and Dunne mechanisms result in variable source areas that expand and contract through wet and dry periods.

The distribution of total global runoff is shown in Figure 11, and the chemical characteristics of this runoff are delineated in Figure 12. It can be seen that total river discharge and the chemistry of the discharge vary from

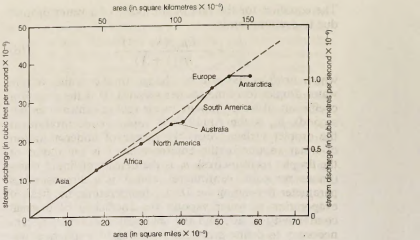


Figure 11: Distribution of total global water discharge to the oceans on a continental basis.

Adapted from *Evolution of Sedimentary Rocks* by Robert M. Garrels and Fred T. Mackenzie; by permission of W.W. Norton & Company, Inc., copyright © 1971 by W.W. Norton & Company, Inc.

continent to continent; some continents are wetter and some drier than the world average, but the deviations are not extreme. The runoff per unit area from Asia and Europe is almost exactly equal to the world average; it is a little lower in Africa and North America; and it is considerably higher in South America. Antarctica is frozen and Australia is arid, and so they contribute little runoff. Also, since their areas are relatively small, they do not affect the global runoff average significantly. The waters draining the continents have quite different chemistries; those from Europe are very rich in calcium and bicarbonates, whereas those from Africa and South America are not. North American and Asian rivers are somewhat intermediate in their concentrations of these dissolved constituents. Such differences in composition reflect a variety of factors, including runoff, temperature, and relief, but certainly the bulk composition of the continental rocks in contact with these waters and their underground sources play a major role (see Figure 3). The surface rocks of Europe are rich in carbonates and those of South America are not; the latter are dominated by sediments rich in silicate minerals.

The chemistry of groundwater and river runoff is being modified by human activities on a global scale. The natural dissolved riverine input of major constituents to the oceans already has been increased by more than 10 percent because of human activities. In the case of sodium, chlorine, and sulfate, the increases are as high as 30 percent. In the United States alone, total water utiliza-

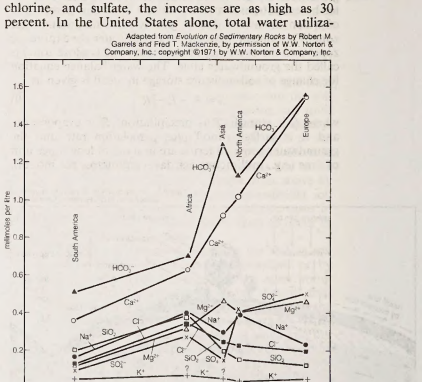


Figure 12: The chemical characteristics of river water draining various continents.

tion is equivalent to one-third of total runoff, with about 2 percent of the water used coming from underground wells. In the southwestern region of the country, water supplies have been tapped heavily and in some areas have been exhausted with no hope of replacement. This extensive utilization of fresh waters in the United States and throughout the globe make them particularly susceptible to pollution. Leachates from fertilizers, herbicides, and pesticides are found in some freshwater bodies; toxic and inorganic or organic chemicals are present; radioactive elements have been detected; and some surface-water bodies have had their salinities increased dramatically, rendering them useless for human consumption. It is therefore imperative that nations closely monitor the utilization of freshwater systems and promote their conservation.

ORIGIN AND EVOLUTION OF THE HYDROSPHERE

It is not very likely that the total amount of water at the Earth's surface has changed significantly over geologic time. Based on the ages of meteorites, the Earth is thought to be 4.6 billion years old. The oldest rocks known date 3.8 billion years in age, and these rocks, though altered by post-depositional processes, show signs of having been deposited in an environment containing water. There is no direct evidence for water for the period between 4.6 and 3.8 billion years ago. Thus, ideas concerning the early history of the hydrosphere are closely linked to theories about the origin of the Earth.

The Earth is thought to have accreted from a cloud of ionized particles around the Sun. This gaseous matter condensed into small particles that coalesced to form a protoplanet, which in turn grew by the gravitational attraction of more particulates. Some of these particles had compositions similar to that of carbonaceous chondrite meteorites, which may contain up to 20 percent water. Heating of this initially cool, unsorted conglomerate by the decay of radioactive elements and the conversion of kinetic and potential energy to heat resulted in the development of the Earth's liquid iron core and the gross internal zonation of the planet (*i.e.,* differentiation into core, mantle, and crust). It has been concluded that the Earth's core formed over a period of about 500 million years. It is likely that core formation resulted in the escape of an original primitive atmosphere and its replacement by one derived from the loss of volatile substances from the planetary interior.

Source of water for the initial hydrosphere

At an early stage the Earth thus did not have water or water vapour at its surface. Once the planet's surface had cooled sufficiently, water contained in the minerals of the accreted material and released at depth could escape to the surface and, instead of being lost to space, cooled and condensed to form the initial hydrosphere. A large, cool Earth most certainly served as a better trap for water than a small, hot body because the lower the temperature, the less likelihood for water vapour to escape, and the larger the Earth, the stronger its gravitational attraction for water vapour. Whether most of the degassing took place during core formation or shortly thereafter or whether there has been significant degassing of the Earth's interior throughout geologic time remains uncertain. It is likely that the hydrosphere attained its present volume early in the Earth's history, and since that time there have been only small losses and gains. Gains would be from continuous degassing of the Earth; the present degassing rate of juvenile water has been determined as being only 0.3 cubic kilometre per year. Water loss in the upper atmosphere is by photodissociation, the breakup of water vapour molecules into hydrogen and oxygen due to the energy of ultraviolet light. The hydrogen is lost to space and the oxygen remains behind. Only about 4.8×10^{-4} cubic kilometre of water vapour is presently destroyed each year by photodissociation. This low rate can be readily explained: the very cold temperatures of the upper atmosphere result in a cold trap at an altitude of about 15 kilometres, where most of the water vapour condenses and returns to lower altitudes, thereby escaping photodissociation. Since the early formation of the hydrosphere, the amount of water vapour in the atmosphere has been regulated by the temperature of the Earth's surface—hence its radiation balance. Higher temperatures imply higher concentrations of atmospheric water vapour, while lower temperatures suggest lower atmospheric levels.

The early hydrosphere. The gases released from the Earth during its early history, including water vapour, have been called excess volatiles because their masses cannot be accounted for simply by rock weathering. An estimate of the excess volatiles is given in Table 6. These volatiles are thought to have formed the early atmosphere of the Earth. At an initial crustal temperature of about 600° C, almost all of these compounds, including H_2O, would have been in the atmosphere. The sequence of events that occurred as the crust cooled is difficult to reconstruct. Below 100° C all of the water would have condensed, and the acid gases would have reacted with the original igneous crustal minerals to form sediments and an initial hydrosphere that was dominated by a salty ocean. If the reaction rates are assumed to have been slow relative to cooling, an atmosphere of 600° C would have contained, together with other compounds, water vapour, carbon dioxide, and hydrogen chloride (HCl) in a ratio of 20:3:1 and cooled to the critical temperature of water (*i.e.,* 374° C). The water therefore would have condensed into an early hot ocean. At this stage, the hydrogen chloride would have dissolved in the ocean (about one mole per litre), but most of the carbon dioxide would have remained in the atmosphere, with only about 0.5 mole per litre in the ocean water. This early acid ocean would have reacted vigorously with crustal minerals, dissolving out silica and cations and creating a residue composed principally of aluminous clay minerals that would form the sediments of the early ocean basins.

Formation of a hot, acid ocean

Table 6: Estimate of Excess Volatiles
(units of 10^{20} grams)

Water	16,660
Total carbon as carbon dioxide	2,500
Chlorine	300
Nitrogen	49
Sulfur	44
Boron, bromine, argon, fluorine, etc.	4

Source: J.C.G. Walker, *Evolution of the Atmosphere* (1977). Macmillan Publishing Co., Inc.

This is one of several possible pathways for the early surface of the Earth. Whatever the actual case, after the Earth's surface had cooled to 100° C, it would have taken only a short time for the remaining acid gases to be consumed in reactions involving igneous rock minerals. The presence of cyanobacteria (*e.g.,* blue-green algae) in the fossil record of rocks older than three billion years attests to the fact that the Earth's surface had cooled to temperatures lower than 100° C by this time, and neutralization of the original acid volatiles had taken place. It is possible, however, that, because of increased greenhouse gas concentrations (see below) in the Early Archean era (about 3.8 to 3.4 billion years ago), the Earth's surface could still have been warmer than today.

If most of the degassing of primary volatile substances from the Earth's interior occurred early, the chloride released by the reaction of hydrochloric acid with rock minerals would be found in the oceans or in evaporite deposits, and the oceans would have a salinity and volume comparable to that of today. This conclusion is based on the assumption that there has been no drastic change in the ratios of volatiles released through geologic time. The overall generalized reaction indicative of the chemistry leading to the formation of the early oceans can be written in the form: primary igneous rock minerals + acid volatiles + $H_2O \rightarrow$ sedimentary rocks + oceans + atmosphere. It should be noted from this equation that, if all the acid volatiles and H_2O were released early in the history of the Earth and in the proportions found today, then the total original sedimentary rock mass-produced would be equal to that of the present, and ocean salinity and volume would be close to those of today as well. If, on the other hand, degassing were linear with time, then the sedimentary rock mass would have accumulated at a linear rate, as would have oceanic vol-

ume. The salinity of the oceans, however, would remain nearly the same if the ratios of volatiles degassed did not change with time. The most likely situation is the one presented here—namely, that major degassing occurred early in Earth's history, after which minor amounts of volatiles were released episodically or continuously for the remainder of geologic time. The salt content of the oceans based on the constant proportions of volatiles released would depend primarily on the ratio of sodium chloride locked up in evaporites to that dissolved in the oceans. If all the sodium chloride in evaporites were added to the oceans today, the salinity would be approximately doubled. This value gives a sense of the maximum salinity that the oceans could have attained throughout geologic time.

Oxygen production

One component absent from the early Earth's surface was free oxygen; it would not have been a constitutent released from the cooling crust. Early production of oxygen was by the photodissociation of water in the Earth's atmosphere, a process that was triggered by the absorption of the Sun's ultraviolet radiation. The reaction is $2H_2O \xrightarrow{h\nu} O_2 + 2H_2$, in which $h\nu$ represents the photon of ultraviolet light. The hydrogen produced would escape into space, while the oxygen would react with the early reduced gases by reactions such as $2H_2S + 3O_2 \rightarrow 2SO_2 + 2H_2O$. Oxygen production by photodissociation gave the early reduced atmosphere a start toward present-day conditions, but it was not until the appearance of photosynthetic organisms approximately three billion years ago that oxygen could accumulate in the Earth's atmosphere at a rate sufficient to give rise to today's oxygenated environment. The photosynthetic reaction leading to oxygen production is given in equation (6).

The transitional hydrosphere. The nature of the rock record from the time of the first sedimentary rocks (approximately 3.8 billion years ago) to about one to two billion years ago suggests that the amount of oxygen in the Earth's atmosphere was significantly lower than it is today and that there were continuous chemical trends in the sedimentary rocks formed and, more subtly, in the composition of the hydrosphere. Figure 13 shows how the chemistry of rocks shifted dramatically during this transitional period. The source rocks of sediments during this time may have been more basaltic than subsequent ones. Sedimentary debris was formed by the alteration of such source rocks in an oxygen-deficient atmosphere and accumulated primarily under anaerobic marine conditions. The chief difference between reactions involving mineral–ocean equilibria at this time and the present day was the role played by ferrous iron (*i.e.*, reduced state of iron). The concentration of dissolved iron in modern oceans is low because of the insolubility of oxidized iron oxides. During the transition stage and earlier, oxygen-deficient environments were prevalent, and these favoured the formation of minerals containing ferrous iron from the alteration of rocks slightly more rich in basalt than those of today. Indeed, iron carbonate siderite and iron silicate greenalite, in close association with chert and iron sulfide pyrite, are characteristic minerals that occur in iron formations of the middle Precambrian (about 2.4 to 1.5 billion years ago). The chert originally was deposited as amorphous silica; equilibrium between amorphous silica, siderite, and greenalite at 25° C and a total pressure of one atmosphere requires a carbon dioxide pressure of about $10^{-2.5}$ atmosphere, or 10 times the present-day value.

Major change in rock chemistry

The oceans of this transitional period can be thought of as a solution that resulted from an acid leach of basaltic rocks, and, because the neutralization of the volatile acid gases was not restricted primarily to land areas as it is today, much of this alteration may have occurred by submarine processes. Anaerobic depositional environments with internal carbon dioxide pressures of about $10^{-2.5}$ atmosphere prevailed, and the oxygen-deficient atmosphere itself may have had a carbon dioxide pressure close to $10^{-2.5}$ atmosphere. If so, the pH of early ocean water was lower than that of modern seawater and the calcium concentration was higher; moreover, the early ocean water was probably saturated with respect to amorphous silica—roughly 120 ppm.

Neutralizing of volatile acid gases

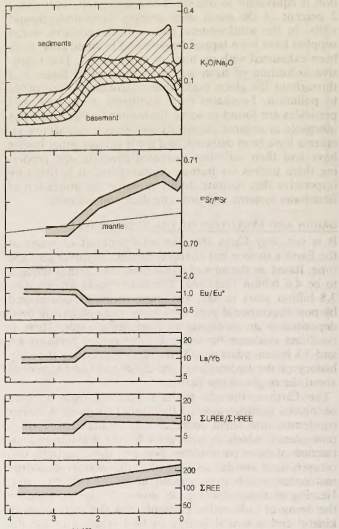


Figure 13: *Variation in the chemistry of rocks as a function of geologic age.* There appears to be a relatively sharp transition in rock composition about two billion years ago. Eu* denotes the nonfractionated value of europium concentration, Σ LREE, Σ HREE, and Σ REE are abbreviations for total light rare earth elements, total heavy rare earth elements, and total rare earth elements, respectively.

After J. Veizer, "The Evolving Exogenic Cycle," in C.B. Gregor *et al.* (eds.), *Chemical Cycles in the Evolution of the Earth*; copyright © 1988, Wiley Interscience, New York City

To simulate what might have occurred, it is helpful to imagine emptying the Pacific basin, throwing in great masses of broken basaltic material, filling it with hydrogen chloride dissolved in water so that the acid becomes neutralized, and then carbonating the solution by bubbling carbon dioxide through it. Oxygen would not be permitted into the system. The hydrochloric acid would leach the rocks, resulting in the release and precipitation of silica and the production of a chloride ocean containing sodium, potassium, calcium, magnesium, aluminum, iron, and reduced sulfur species in the proportions present in the rocks. As complete neutralization was approached, the aluminum could begin to precipitate as hydroxides and then combine with precipitated silica to form cation-deficient aluminosilicates. As the neutralization process reached its end, the aluminosilicates would combine with more silica and with cations to form such minerals as chlorite, and ferrous iron would combine with silica and sulfur to produce greenalite and pyrite. In the final solution, chlorine would be balanced by sodium and calcium in roughly equal proportions, with subordinate amounts of potassium and magnesium; aluminum would be quantitatively removed, and silicon would be at saturation with amorphous silica. If this solution were then carbonated, calcium would be removed as calcium carbonate, and the chlorine balance would be maintained by abstraction of more sodium from the primary rock. The sediments formed in this system would contain chiefly silica, ferrous iron silicates, chloritic minerals, calcium carbonate, calcium-magnesium carbonates, and small amounts of pyrite.

If the hydrogen chloride added were in excess of the carbon dioxide, the resultant oceans would have a high content of calcium chloride ($CaCl_2$), but with a pH still near neutrality. If the carbon dioxide added were in excess of the chlorine, calcium would be precipitated as carbonate until it reached a level roughly that of present-day ocean waters—namely, a few hundred parts per million.

If this newly created ocean were left undisturbed for several hundred million years, its waters would evaporate and be transported onto the continents (in the form of precipitation); streams would transport their loads into it. The sediments produced in this ocean would be uplifted and incorporated into the continents. The influence of the continental debris would gradually be felt and the pH might change somewhat. Iron would be oxidized out of the ferrous silicates to yield iron oxides, but the composition of the water would not vary substantially.

The primary minerals of igneous rocks are all mildly basic compounds. When these minerals react in excess with acids such as hydrogen chloride and carbon dioxide, they produce neutral or mildly alkaline solutions as well as a set of altered aluminosilicate and carbonate reaction products. It is improbable that seawater has changed through time from a solution approximately in equilibrium with these reaction products—*i.e.*, with clay minerals and carbonates.

Little change in chemical and mineralogical composition

The modern hydrosphere. It is likely that the hydrosphere achieved its modern chemical characteristics about 1.5 to two billion years ago. The chemical and mineralogical compositions and the relative proportions of sedimentary rocks of this age differ little from their counterparts of the Paleozoic era (from 540 to 245 million years ago). Calcium sulfate deposits of late Precambrian age (about 1.5 billion to 540 million years ago) attest to the fact that the acid sulfur gases had been neutralized to sulfate by this time. Chemically precipitated ferric oxides in late Precambrian sedimentary rocks indicate available free oxygen, whatever its percentage. The chemistry and mineralogy of middle and late Precambrian shales are similar to those of Paleozoic shales. The carbon isotopic signature of carbonate rocks has been remarkably constant for more than three billion years, indicating exceptional stability in size and fluxes related to organic carbon. The sulfur isotopic signature of sulfur phases in rocks strongly suggests that the sulfur cycle involving heterotrophic bacterial reduction of sulfate was in operation 2.7 billion years ago. It therefore appears that continuous cycling of sediments similar to those of today has occurred for 1.5 to two billion years and that these sediments have controlled hydrospheric, and particularly oceanic, composition.

It was once thought that the saltiness of the modern oceans simply represents the storage of salts derived from rock weathering and transported to the oceans by fluvial processes. With increasing knowledge of the age of the Earth, however, it was soon realized that, at the present rate of delivery of salts to the ocean or even at much reduced rates, the total salt content and the mass of individual salts in the oceans could be attained in geologically short time intervals compared to the planet's age. The total mass of salt in the oceans can be accounted for at today's rates of stream delivery in about 12 million years. The mass of dissolved silica in ocean water can be doubled in just 20,000 years by the addition of stream-derived silica; to double the sodium content would take 70 million years. It then became apparent that the oceans were not merely an accumulator of salts; rather, as water evaporated from the oceans, together with some salt, the salts introduced must be removed in the form of minerals deposited in sediments. Accordingly, the concept of the oceans as a chemical system changed from that of a simple accumulator to that of a steady-state system in which rates of inflow of materials equal rates of outflow. The steady-state concept permits influx to vary with time, but the inflow would be matched by nearly simultaneous and equal variation of efflux.

The oceans as a steady-state system

In recent years, this steady-state conceptual view of the oceans has undergone some modification. In particular, it has been found necessary to treat components of ocean water in terms of all their influxes and effluxes and to be more cognizant of the time scale of application of the steady-state concept. Indeed, the recent increase in the carbon dioxide concentration of the atmosphere due to the burning of fossil fuels may induce a change in the pH and dissolved inorganic carbon concentrations of surface ocean water on a time scale measured in hundreds of years. If fossil-fuel burning were to cease, return to the original state of seawater composition could take thousands of years. Ocean water is not in steady state with respect to carbon on these time scales, but on a longer geologic time scale it certainly could be. Even on this longer time scale, however, oceanic composition has varied because of natural changes in the carbon dioxide level of the atmosphere and because of other factors. Figure 14 shows a model calculation of how the carbon dioxide pressure of the atmosphere and the chemistry of seawater may have varied for the past 100 million years.

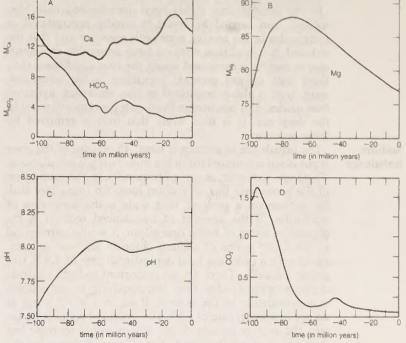


Figure 14: (A,B,C) Results of a model calculation showing possible changes in some parameters of ocean water chemistry and (D) atmospheric carbon dioxide concentration for the past 100 million years. Molar concentration is denoted by M.

It appears that the best description of modern seawater composition is that of a chemical system in a dynamic quasi-steady state. Changes in composition may occur over time, but the system always seems to return to a time-averaged, steady-state composition. In other words, since 1.5 to two billion years ago, evolutionary chemical changes in the hydrosphere have been small when viewed against the magnitude of previous change.

Table 7 shows a tentative mass balance for the modern oceans in terms of sources and sinks of some important elements. It should be noted that rivers supply dissolved constituents to the oceans, whereas high- and low-temperature reactions between seawater and submarine basalts and reactions in sediment pore waters may add or remove constituents from ocean water. Biological processes involved in the formation of the opaline silica skeletons of diatoms and radiolarians and the carbonate skeletons of planktonic foraminiferans and coccolithophorids chiefly remove calcium and silica from seawater. Exchange reactions between river-borne clays entering seawater are particularly significant for sodium and calcium ions. The mass balance of Table 7, highlighting the major processes that govern the steady-state composition of seawater on a long time scale for the elements shown, is reasonable except for sodium and iron. Most of the carbon imbalance represents carbon released to the ocean–atmosphere system during precipitation of carbonate minerals—*i.e.*,

$$Ca^{2+} + 2HCO_3^- = CaCO_3 + CO_2 + H_2O. \qquad (11)$$

Table 7: Tentative Mass Balance of Some Dissolved Species in Seawater (in units of 10^{12} grams per year)

element	river input*	hydrothermal reactions	submarine weathering	pore water	exchange reactions	biological removal	net balance
Na	+131†	−11‡	−5	−9(?)	−43		+63
Mg	+129	−60	+26	−82	−8	−6	−1
Al	+1.9				0	−2	0
Si	+203	+25	+10	+185		−375	+48
K	+50	−5	−10	−26	−8		+1
Ca	+495	+72	+32	+40	+38	−645	+32
Fe	+1.5	+15	+7	0	0	0	+23.5
C§	+389	+60	+60	+60	+5	−215	+299§

*Corrected for cyclic salts. †The + indicates input to the ocean. ‡The − indicates output from the ocean. §The values provided here are for dissolved inorganic carbon. §215×10^{12} grams of C per year of this net imbalance for carbon represents primarily carbon released to the ocean–atmosphere system during precipitation of carbonate minerals.

Source: Adapted from R.W. Wollast and F.T. Mackenzie, "Global Cycle of Silica," in S.R. Aston (ed.), *Silicon Geochemistry and Biogeochemistry* (1983). D. Reidel Publishing Co.

In the case of iron, it has been documented that "dissolved" iron carried by rivers is rapidly precipitated as hydroxides in the mixing zone with seawater and that the reduced dissolved iron released from anaerobic sediments also is rapidly precipitated under the oxic conditions (*i.e.,* those with oxygen present) prevailing in the water column. Iron is also precipitated as iron smectites, hydrated iron oxides, and nontronite (iron-rich montmorillonite) in the deep sea. It is thus likely that iron is removed by these processes.

Sodium imbalance

The imbalance in sodium is large; 45 percent of the river input is not accounted for in the mass balance calculations. There are, however, major uncertainties in the estimation of the pore-water flux of sodium ions. An important sink for sodium on a geologic time scale is the formation of evaporites. If the amount of unbalanced sodium is expressed in terms of halite deposition, it would correspond to 1.6×10^{14} grams of sodium chloride per year as compared with a potential total depositional rate of 3.3×10^{14} grams annually. There are no important sodium chloride deposits forming today; thus, one possibility is that sodium is accumulating in the oceans. If so, in 6×10^6 years at an accumulation rate of 63×10^{12} grams of sodium annually, the average salinity of the oceans would increase less than one part per thousand. The chlorine balance for the oceans, however, indicates that it is likely that the major problem in the imbalance for sodium lies in the flux estimates for sediment pore waters and perhaps submarine weathering processes.

Modern seawater chemistry has been characteristic of roughly the past 600 million years of ocean history. Evaporite sediments provide strong evidence that the composition of seawater has not varied a great deal during this interval of geologic time. Nonetheless, it seems likely, as shown in Figure 14, that fluctuations did occur, particularly in the concentrations of calcium, magnesium, and sulfate ions. The isotopic composition of sulfur in seawater, as recorded in evaporites, has varied dramatically during the past one billion years. Although it is difficult to relate these isotopic fluctuations to the calcium and sulfate concentrations of seawater, some scientists believe that the fluctuations do in fact imply changes in the latter. Furthermore, the major features of the sulfur isotopic curve for evaporites versus Phanerozoic time is similar to that of the strontium-87/strontium-86 ratio, and perhaps the strontium/calcium ratio, of sedimentary materials during this time interval. Such covariation is consistent with a model in which fluxes related to alteration of seafloor basalts and continental river runoff vary with time, resulting in variation in seawater composition.

Changes in the chemistry of the atmosphere–hydrosphere

The chemistry of the atmosphere has certainly changed significantly during the past one billion years of Earth history. A modification of this kind implies changes in the chemistry of the hydrosphere as well. Oxygen in the atmosphere rose substantially between two billion years ago and the beginning of the Phanerozoic eon (*i.e.,* 540 million years ago), whereas atmospheric carbon dioxide levels probably decreased. This change led in general to a progressively more oxygenated and less acidic hydrosphere. It is likely that the development of higher land plants during the Devonian period (from 408 to 360 million years ago) resulted in an increase in atmospheric oxygen and a decrease in carbon dioxide. Air trapped within bubbles in Arctic and Antarctic ice shows that the carbon dioxide content of the atmosphere during the climax of the last ice age was about 180 parts per million by volume (ppmv), and atmospheric CO_2 levels reached approximately 280 ppmv during the last great interglacial of 120,000 years ago, long before modern society initiated its extensive fossil-fuel burning and deforestation activities (see below *Buildup of greenhouse gases*). These atmospheric changes in themselves can influence the chemistry of the hydrosphere, but they also appear to be coupled with other changes in the rock–ocean–atmosphere–biota system that strongly affect hydrospheric chemistry. For example, though surface waters probably remained oxygenated during the Cretaceous and Devonian periods of Earth history, there is evidence that intermediate and deep ocean waters were more anoxic (oxygen depleted) than today. These "anoxic events" are characterized by dramatic changes in the Earth's ocean–atmosphere system, including changes in the rates of the cyclic transfer of elements at the Earth's surface and in atmospheric composition. The probable impact of a bolide (either an asteroid or comet) and increased volcanism at the end of the Cretaceous (about 66.4 million years ago), though still the subject of hot debate, certainly could have caused short-term changes in the chemistry of the Earth's atmosphere-hydrosphere. Scientists speculate that such an event could have had any of several results, including (1) changes in the Earth's radiant-energy balance because of vast amounts of particulate and gaseous input into the atmosphere and subsequent cooling, (2) acid rain stemming from the input into the atmosphere of nitrogen gases generated by the bolide impact, (3) increased trace metal fluxes to the hydrosphere brought on by the destruction of the bolide and increased volcanism, and perhaps (4) increased "anoxia" of the hydrosphere owing to the death of land and marine organisms.

Whatever the case, evidence of change is present in the rock record albeit the composition of the modern hydrosphere has not varied greatly for the past one billion years. Moreover, as will be seen in the following section, humankind is modifying not only the chemistry of local and regional water bodies but also that of the entire global atmosphere-hydrosphere by increasing the rates of input of natural substances and by introducing new synthetic substances to the environment.

IMPACT OF HUMAN ACTIVITIES ON THE HYDROSPHERE

The activities of modern society are having a severe impact on the hydrologic cycle. The dynamic steady state is being disturbed by the discharge of toxic chemicals, radioactive substances, and other industrial wastes and by the seepage of mineral fertilizers, herbicides, and pesticides into surface and subsurface aquatic systems. Inadvertent and deliberate discharge of petroleum, improper sewage

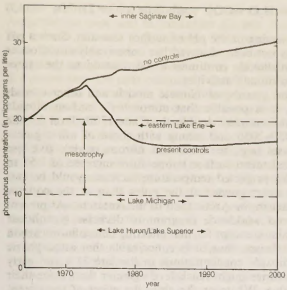


Figure 15: *Past and calculated future trends of phosphorus concentrations in the waters of Lake Ontario.* It should be noted how the present controls can bring this lake water into a state of mesotrophy (between eutrophy and oligotrophy; thus characterized by a moderate amount of dissolved nutrients). Other Great Lakes water compositions are also shown.

After S.C. Chapra, "Simulation of Recent and Projected Total Phosphorus Trends in Lake Ontario," *Journal of Great Lakes Research*, vol. 6 (1980)

disposal, and thermal pollution also are seriously affecting the quality of the hydrosphere.

The present discussion focuses on three major problems—eutrophication, acid rain, and the buildup of the so-called greenhouse gases. Each exemplifies human interference in the hydrologic cycle and its far-reaching effects.

Eutrophication. Historically, aquatic systems have been classified as oligotrophic or eutrophic. Oligotrophic waters are poorly fed by the nutrients nitrogen and phosphorus and have low concentrations of these constituents. There is thus low production of organic matter by photosynthesis (equation [6]) in such waters. By contrast, eutrophic waters are well supplied with nutrients and generally have high concentrations of nitrogen and phosphorus and, correspondingly, large concentrations of plankton owing to high biological productivity. The waters of such aquatic systems are usually murky, and lakes and coastal marine systems may be oxygen-depleted at depth. The process of eutrophication is defined as high biological productivity resulting from increased input of nutrients or organic matter into aquatic systems. For lakes, this increased biological productivity usually leads to decreased lake volume because of the accumulation of organic detritus. Natural eutrophication occurs as aquatic systems fill in with organic matter; it is distinct from cultural eutrophication, which is caused by human intervention. The latter is characteristic of aquatic systems that have been artificially enriched by excess nutrients and organic matter from sewage, agriculture, and industry. Naturally eutrophic lakes may produce 75–250 grams of carbon per square metre per year, whereas those lakes experiencing eutrophication because of human activities can support 75–750 grams per square metre per year. Commonly, culturally eutrophic aquatic systems may exhibit extremely low oxygen concentrations in bottom waters. This is particularly true of stratified systems, as, for instance, lakes during summer where concentrations of molecular oxygen may reach levels of less than about one milligram per litre—a threshold for various biological and chemical processes.

Cultural eutrophication

Aquatic systems may change from oligotrophic to eutrophic, or the rate of eutrophication of a natural eutrophic system may be accelerated by the addition of nutrients and organic matter due to human activities. The process of cultural eutrophication, however, can be reversed if the excess nutrient and organic matter supply is shut off. This response is demonstrated in Figure 15, which shows how the Great Lakes of the north central United States have responded to government regulations of the early 1970s aimed at reducing phosphorus inputs (principally from detergents) into these lakes.

Not only do freshwater aquatic systems undergo cultural eutrophication, but coastal marine systems also may be affected by this process. On a global scale, the input by rivers of organic matter to the oceans today is twice the input in prehuman times, and the flux of nitrogen, together with that of phosphorus, has more than doubled. This excess loading of carbon, nitrogen, and phosphorus is leading to cultural eutrophication of marine systems. In several polluted eastern U.S. estuaries and in some estuaries of western Europe (*e.g.*, the Scheldt of Belgium and The Netherlands), all of the dissolved silica brought into the estuarine waters by rivers is removed by phytoplankton growth (primarily diatoms) resulting from excess fluxes of nutrients and organic matter. In the North Sea, there is now a deficiency of silica and an excess of nitrogen and phosphorus, which in term has led to a decrease in diatom productivity and an increase in cyanobacteria productivity—a biotic change brought about by cultural eutrophication.

Acid rain. The emission of sulfur dioxide and nitrogen oxides to the atmosphere by human activities—primarily fossil-fuel burning—has led to the acidification of rain and freshwater aquatic systems. Acid rain is a worldwide problem and has been well documented for the eastern United States and the countries of western Europe.

Acid rain is defined as precipitation with a pH of less than 5.7 that results from reactions involving gases other than carbon dioxide. The overall reactions that produce such precipitation are those of equations (1) through (3) and

$$NO_2 + OH \rightarrow HNO_3, \qquad (12)$$

followed by dissociation of the HNO_3 to $H^+ + NO_3^-$.

Figure 16 shows the average $pH = -\log a_{H^+}$ (a_{H^+} is activity of the hydrogen ion) of precipitation over the eastern United States for the period October 1979 through September 1980. The low pH values are a result of equilibration of rainwater with the atmospheric acid gases of carbon, nitrogen, and sulfur. Equilibration only with atmospheric carbon dioxide would give a pH of 5.7. The significantly lower values are a result of reactions with nitrogen- and sulfur-bearing gaseous atmospheric components derived primarily from fossil-fuel burning sources. Nitrate and sulfate concentrations in precipitation over the eastern United States are strongly correlated with pH—the lower the pH of rain, the higher the concentrations of nitrate and sulfate. Such low pH values and increased

Reproduced, with permission, from W.D. Bischoff, V.L. Paterson, and F.T. Mackenzie, "Geochemical Mass Balance for Sulfur- and Nitrogen-Bearing Acid Components: Eastern United States," in D.P. Bricker (ed.), *Geological Aspects of Acid Deposition* (1984), Butterworth Publishers, Stoneham, Mass.

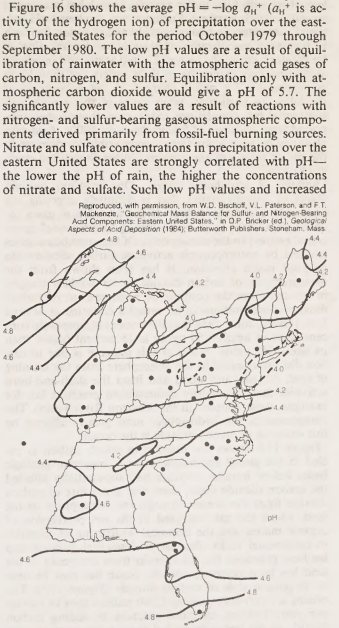


Figure 16: The average pH of precipitation during the period October 1979 to September 1980 in the eastern United States.

Table 8: Estimated Annual Fluxes of the Acid-Precipitation Constituents Sulfur and Nitrogen for the Eastern United States

flux	sulfur (10^{11} grams)	nitrogen (10^{11} grams)
A. Emission	113.9	38.8
B. Wet deposition	21.4	7.6
C. Present river	128.0	10.9
D. Pre-human river	76.8	1.4
E. Anthropogenic river (C–D)	51.2	9.5
F. Dry deposition (B–E)	29.8	1.9
G. Atmospheric transport (A–E)	62.7	29.3
H. Atmospheric transport to Canada (0.52 × grams for sulfur) (0.45 × grams for nitrogen)	32.6	13.2
I. Atmospheric transport to Atlantic (0.48 × grams for sulfur) (0.55 × grams for nitrogen)	30.1	16.1

Source: W.D. Bischoff, V.L. Paterson, and F.T. Mackenzie, "Geochemical Mass Balance for Sulfur- and Nitrogen-Bearing Acid Components: Eastern United States," in O.P. Bricker (ed.), *Geological Aspects of Acid Deposition* (1984). Butterworth Publishers.

nitrate and sulfate concentrations also are found in the rains of western Europe and other industrialized regions of the world.

Table 8 provides an illustrative example of events occurring globally—namely, the processes—that remove anthropogenic emissions of sulfur dioxide and nitrogen oxides from the atmosphere of the eastern United States. Wet and dry deposition also removes the hydrogen ion produced in the rain by the oxidation and hydrolysis of these acid gases. This excess hydrogen ion can bring about the acidification of freshwater aquatic systems, particularly those with little buffer capacity (*e.g.*, lakes situated in crystalline rock terrains). Furthermore, the lower pH values of rainwater, and consequently of soil water, can lead to increased mobilization of aluminum. Acidification of freshwater lakes in the eastern United States and increased aluminum concentrations in their waters are thought to be responsible for major changes in the ecosystems of the lakes. In particular, many lakes of this region lack substantial fish populations today, even though they supported large numbers of fish in the early 1900s. Acid rain also may be among the factors responsible for damage to the major forests of the eastern United States and western Europe.

Impact on lake ecosystems

Buildup of greenhouse gases. One problem brought about by human action that is definitely affecting the hydrosphere globally is that of the greenhouse gases (so called because of their heat-trapping "greenhouse" properties) emitted to the atmosphere. Of the greenhouse gases released by anthropogenic activities, carbon dioxide has received much attention. It has been shown from the measurements of carbon dioxide in air bubbles trapped in ice and from the continuous measurement of carbon dioxide concentrations in air samples collected at Mauna Loa, Hawaii, since 1958 that the present atmospheric concentration of nearly 350 ppmv is 25 percent higher than its late-1700s value. Much of this increase is due to carbon dioxide released to the atmosphere from the burning of coal, oil, gas, and wood and from the slash-and-burn activities that accompany deforestation practices (as, for example, those adopted in the Amazon River basin). The component of the hydrosphere most greatly affected by this emission of carbon dioxide is the ocean.

Figure 17A shows the manner in which carbon is cycled in the global environment on a long-term geologic basis. Before human activities had substantially affected the carbon dioxide cycle, there was a net flux of carbon dioxide from the oceans through the atmosphere to the land, where the gas was used in the net production of organic matter and the chemical weathering of minerals in continental rocks. Because of fossil-fuel burning and land-use practices, the net transfer from the ocean to the land has been reversed, and the ocean has now become an important sink of carbon dioxide (Figure 17B). The oceans are currently gaining 2,340 million tons of carbon per year. The net chemical reaction of adding carbon dioxide to the ocean (provided there is no reaction with carbonate solids) is

$$CO_2 + H_2O + CO_3^{2-} \text{ (carbonate ion)} = 2HCO_3^- \quad (13)$$

and a lowering of the pH of surface seawater. Such a pH effect has not been observed but conceivably could occur if carbon dioxide continues to be released to the atmosphere by human activities.

Based on greenhouse climate models and other considerations, it is possible that atmospheric carbon dioxide concentrations may double their late-1700s level by the years 2030–2050 and, along with those of other greenhouse gases (*e.g.*, methane and nitrous oxide), give rise to a global mean surface temperature increase of 1.5° to 5° C. This projected temperature increase would be two to three times greater at the poles than at the equator and greater in the Arctic than in the Antarctic. At present there is no worldwide program to decrease greenhouse gas emissions, except for that affecting chlorofluorocarbon (Freon) releases; thus, it is conceivable that atmospheric carbon dioxide concentrations in the late 21st and early 22nd centuries might reach levels greater than twice their 1700s value. Whatever the case, the effect of the potential rise in surface temperature would be to speed up the hydrologic cycle and probably the rate of chemical weathering of continental rocks. Increases of 4 to 7 percent in the global mean evaporation and precipitation rates might occur for a doubling of the carbon dioxide level and a few degrees rise in global mean temperature. The effect on the water balance would be regional in nature, with some places becoming wetter and others drier. In general, there would be a trend toward greater and longer periods of summer dryness induced by lower soil moisture content and higher evaporation rates in the mid-latitudes of the Northern Hemisphere. In the arid western regions of the United States, which depend on irrigation for growing plants, severe water shortages could occur. By contrast, precipitation and runoff might increase, except in summer, at latitudes beyond 60° N because of a greater poleward transport of moisture. In summer, in a zone centred around 60° N, greater dryness might occur due to an earlier end of snowmelt.

Global warming and its potential effects on the hydrologic cycle

Global warming could further affect the hydrologic cycle

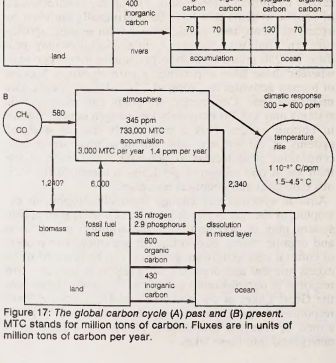


Figure 17: *The global carbon cycle* (A) *past and* (B) *present.* MTC stands for million tons of carbon. Fluxes are in units of million tons of carbon per year.

by the melting of ice and snow in the Greenland and Antarctic ice caps and in mountain glaciers, resulting in the transfer of water to the oceans. This process, together with thermal expansion of the oceans because of global warming, could lead to a slow rise in sea level of about 0.7 metre over the next century. If the West Antarctic ice sheet were to disintegrate, a much larger and more rapid rise in sea level of 5–6 metres could occur over the next several hundred years. The melting of all glacial ice would raise the sea level about 56 metres. It is also possible that a global warming could result in a reduction in the areal extent and thickness of sea ice in the Arctic and circum-Antarctic regions. Complete melting of the Arctic sea ice might occur, causing a northward shift in storm tracks and a reduction in Northern Hemispheric precipitation during the spring and fall. Furthermore, a worldwide reduction in sea ice might lead to increased evaporation from the ocean and increased low-altitude cloudiness, which would reflect solar radiation and cause cooling.

The potential changes in the hydrologic cycle induced by a global warming resulting from anthropogenic emissions of greenhouse gases do not seem great. Yet, their consequences could be severe for ecosystems and human populations, especially since the latter are so sensitive to and dependent on such changes. A global rise in sea level of one metre, for example, would almost completely inundate the coastal areas of Bangladesh. Island nations and continental beaches and cities would be endangered. Agricultural lands could be displaced, just as patterns of arid, semiarid, and wet lands might become modified. It is essential that society plan for such potential changes so that, if they do occur, appropriate adjustments can be made to accommodate them.

BIBLIOGRAPHY. General introductory discusssions on the distribution of water on and around the Earth and the role of water in supporting life are found in CYNTHIA A. HUNT and ROBERT M. GARRELS, *Water: The Web of Life* (1972); C.L. MANTELL and A.M. MANTELL, *Our Fragile Water Planet: An Introduction to the Earth Sciences* (1976); ELIZABETH KAY BERNER and ROBERT A. BERNER, *The Global Water Cycle: Geochemistry and Environment* (1987); H.M. RAGHUNATH, *Ground Water*, 2nd ed. (1987); EBERHARD CZAYA, *Rivers of the World* (1981; originally published in German, 1981); MARY J. BURGIS and PAT MORRIS, *The Natural History of Lakes* (1987); and NEIL WELLS, *The Atmosphere and Ocean: A Physical Introduction* (1986).

Biogeochemical properties of the hydrosphere are discussed in RONALD J. GIBBS, "Mechanisms Controlling World Water Chemistry," *Science* 170(3962):1088–1090 (1970); R. WOLLAST and FRED T. MACKENZIE, "Global Cycle of Silica," in S.R. ASTON (ed.), *Silicon Geochemistry and Biogeochemistry* (1983), pp.39–76; P. BUAT-MENARD, "Particle Geochemistry in the Atmosphere and Oceans," in PETER S. LISS and W. GEORGE N. SLINN (eds.), *Air-Sea Exchange of Gases and Particles* (1983), pp. 455–532; ROBERT A. BERNER, A.C. LASAGA, and ROBERT M. GARRELS, "The Carbonate-Silicate Geochemical Cycle and Its Effect on Atmospheric Carbon Dioxide Over the Past 100 Million Years," *American Journal of Science* 283(7):641–683 (1983); LAWRENCE A. HARDIE and HANS P. EUGSTER, "The Evolution of Closed-Basin Brines," *Mineralogical Society of America Special Paper* 3:273–290 (1970); JAMES I. DREVER, *The Geochemistry of Natural Waters* (1982); A. LERMAN, *Geochemical Processes: Water and Sediment Environments* (1979, reprinted 1988); G. EVELYN HUTCHINSON, *A Treatise on Limnology*, vol. 1 (1975); WERNER STUMM (ed.), *Chemical Processes in Lakes* (1985); and GEORG MATTHESS, *The Properties of Groundwater* (1982; originally published in German, 1973).

Analyses of the processes of the hydrologic cycle and its utilization are presented in ROBERT C. AVERETT and DIANE M. MCKNIGHT (eds.), *Chemical Quality of Water and the Hydrologic Cycle* (1987); R.A. FREEZE, "A Stochastic-Conceptual Analysis of Rainfall-Runoff Processes on a Hillslope," *Water Resources Research* 16(2):391–408 (1980); T. DUNNE, "Field Studies of Hillslope Flow Processes," in M.J. KIRKBY (ed.), *Hillslope Hydrology* (1978), pp. 227–293; RAY K. LINSLEY and JOSEPH B. FRANZINI, *Water-Resources Engineering*, 3rd ed. (1979); MARK J. HAMMER and KENNETH A. MACKICHAN, *Hydrology and Quality of Water Resources* (1981); and ALVIN S. GOODMAN, *Principles of Water Resources Planning* (1984).

For the evolution of the hydrosphere, see JAMES C.G. WALKER, *Evolution of the Atmosphere* (1977); J. VEIZER, "The Evolving Exogenic Cycle," in C. BRYAN GREGOR *et al.* (eds.), *Chemical Cycles in the Evolution of the Earth* (1988), pp. 175–220; HEINRICH D. HOLLAND, *The Chemical Evolution of the Atmosphere and Oceans* (1984); and ROBERT M. GARRELS and FRED T. MACKENZIE, *Evolution of Sedimentary Rocks* (1971).

The impact of human activities on the hydrosphere is studied in ROBERT M. GARRELS, FRED T. MACKENZIE, and CYNTHIA A. HUNT, *Chemical Cycles and the Global Environment: Assessing Human Influences* (1975); ARTHUR N. STRAHLER and ALAN H. STRAHLER, *Environmental Geoscience: Interaction Between Natural Systems and Man* (1973); W.D. BISCHOFF, V.L. PATERSON, and FRED T. MACKENZIE, "Geochemical Mass Balance for Sulfur- and Nitrogen-Bearing Acid Components: Eastern United States," in OWEN P. BRICKER (ed.), *Geological Aspects of Acid Deposition* (1984), pp. 1–21; S.C. CHAPRA, "Simulation of Recent and Projected Total Phosphorus Trends in Lake Ontario," *Journal of Great Lakes Research* 6(2):101–112 (1980); BRIAN HENDERSON-SELLERS and H.R. MARKLAND, *Decaying Lakes: The Origins and Control of Cultural Eutrophication* (1987); THOMAS D. BROCK, *A Eutrophic Lake: Lake Mendota, Wisconsin* (1985); G. DENNIS COOKE *et al.*, *Lake and Reservoir Restoration* (1986); SVEN OLOF RYDING and WALTER RAST (eds.), *The Control of Eutrophication of Lakes and Reservoirs* (1989); LOUIS THIBODEAUX, *Chemodynamics, Environmental Movement of Chemicals in Air, Water, and Soil* (1979); STANLEY E. MANAHAN, *Environmental Chemistry*, 4th ed. (1984); HOWARD S. PEAVY, DONALD R. ROWE, and GEORGE TCHOBANOGLOUS, *Environmental Engineering* (1985); JAMES L. REGENS and ROBERT W. RYCROFT, *The Acid Rain Controversy* (1988); and DANIEL D. CHIRAS, *Environmental Science: A Framework for Decision Making*, 2nd ed. (1988).

(F.T.M.)

Ice and Ice Formations

Ice occurs on the Earth's continents and surface waters in a variety of forms. Most notable are the continental glaciers (ice sheets) that cover much of Antarctica and Greenland. Smaller masses of perennial ice called ice caps occupy parts of Arctic Canada and other high-latitude regions, and mountain glaciers occur in more restricted areas, such as mountain valleys and the flatlands below. Other occurrences of ice on land include the different types of ground ice associated with permafrost—that is, permanently frozen soil common to very cold regions. In the oceanic waters of the polar regions, icebergs occur when large masses of ice break off from glaciers or ice shelves and drift away. The freezing of seawater in these regions results in the formation of sheets of sea ice known as pack ice. During the winter months similar ice bodies form on lakes and rivers in many parts of the world. The origins, characteristics, and distribution of all such ice formations are treated in this article, as are the structure and properties of ice in general. For a detailed account of the widespread occurrences of glacial ice during the Earth's past, see the articles GEOCHRONOLOGY and CLIMATE AND WEATHER. See also GEOMORPHIC PROCESSES and CONTINENTAL LANDFORMS for the effects of glaciation.

This article is divided into the following sections:

Structure and properties of ice 732
- Structure 732
  - The water molecule
  - The ice crystal
- Properties 733
  - Mechanical properties
  - Thermal properties
  - Optical properties
  - Electromagnetic properties

Ice in lakes and rivers 734
- Geographic extent
- The seasonal cycle
- Ice in lakes 734
  - Ice formation
  - Ice growth
  - Ice decay
  - Geographic distribution
- Ice in rivers 736
  - Formation and growth
  - Decay and ice jams
  - Ice modification
  - Geographic distribution

Glaciers and ice sheets 738
- General observations 738
  - Main types of glaciers
  - Distribution of glaciers
  - Glaciers and climate
- Formation and characteristics of glacier ice 739
  - Transformation of snow to ice
  - Mass balance
  - Heat or energy balance
  - Glacier flow
  - Response of glaciers to climatic change
  - Glaciers and sea level
- The great ice sheets 741
  - Antarctic Ice Sheet
  - Greenland Ice Sheet
  - Mass balance of the ice sheets
  - Flow of the ice sheets
  - Information from deep cores
- Mountain glaciers 744
  - Classification of mountain glaciers
  - Surface features
  - Mass balance of mountain glaciers
  - Flow of mountain glaciers
  - Glacier hydrology
  - Glacier floods
  - Glacier surges
  - Tidewater glaciers

Icebergs and pack ice 747
- Icebergs 747
  - Formation and distribution
  - Size of icebergs
  - Age and melting
  - Wind and current effects
  - Sediment transport
  - Iceberg detection and destruction
- Pack ice 751
  - Formation and characteristics
  - Arctic pack ice
  - Antarctic pack ice
  - Sea ice forecasting and reconnaissance

Permafrost 752
- Distribution in the Northern Hemisphere 753
  - Permafrost zones
  - Study of permafrost
- Origin and stability of permafrost 754
  - Air temperature and ground temperature
  - Climatic change
- Local thickness 755
  - Effects of climate
  - Effects of water bodies
  - Effects of solar radiation, vegetation, and snow cover
- Ice content 755
  - Types of ground ice
  - Ice wedges
- Surface manifestations of permafrost and seasonally frozen ground 757
  - Areas underlain by permafrost
  - Features related to seasonal frost
- Problems posed by permafrost 758
  - Permafrost engineering
  - Development in permafrost areas

Bibliography 759

Structure and properties of ice

STRUCTURE

The water molecule. Ice is the solid state of water, a normally liquid substance that freezes to the solid state at temperatures of 0° C (32° F) or lower and expands to the gaseous state at temperatures of 100° C (212° F) or higher. Water is an extraordinary substance, anomalous in nearly all its physical and chemical properties and easily the most complex of all the familiar substances that are single-chemical compounds. Consisting of two atoms of hydrogen (H) and one atom of oxygen (O), the water molecule has the chemical formula H_2O. These three atoms are covalently bonded (*i.e.*, their nuclei are linked by attraction to shared electrons) and form a specific structure, with the oxygen atom located between the two hydrogen atoms. The three atoms do not lie in a straight line, however. Instead, the hydrogen atoms are bent toward each other, forming an angle of about 105°.

The three-dimensional structure of the water molecule can be pictured as a tetrahedron with an oxygen nucleus centre and four legs of high electron probability, as shown in Figure 1. The two legs in which the hydrogen nuclei are present are called bonding orbitals. Opposite the bonding orbitals and directed to the opposite corners of the tetrahedron are two legs of negative electrical charge. Known as the lone-pair orbitals, these are the keys to water's peculiar behaviour, in that they attract the hydrogen nuclei of adjacent water molecules to form what are called hydrogen bonds. These bonds are not especially strong, but, because they orient the water molecules into a specific configuration, they significantly affect the properties of water in its solid, liquid, and gaseous states.

Hydrogen bonds

In the liquid state, most water molecules are associated

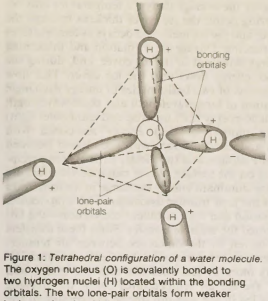


Figure 1: *Tetrahedral configuration of a water molecule.* The oxygen nucleus (O) is covalently bonded to two hydrogen nuclei (H) located within the bonding orbitals. The two lone-pair orbitals form weaker hydrogen bonds with hydrogen nuclei in neighbouring water molecules.

From V.F. Petrenko, *Structure of Ordinary Ice I_h, Part 1: Ideal Structure of Ice*, October 1993. Cold Regions Research & Engineering Laboratory, U.S. Army Corps of Engineers

in a polymeric structure—that is, chains of molecules connected by weak hydrogen bonds. Under the influence of thermal agitation, there is a constant breaking and reforming of these bonds. In the gaseous state, whether steam or water vapour, water molecules are largely independent of one another, and, apart from collisions, interactions between them are slight. Gaseous water, then, is largely monomeric—*i.e.*, consisting of single molecules—although there occasionally occur dimers (a union of two molecules) and even some trimers (a combination of three molecules). In the solid state, at the other extreme, water molecules interact with one another strongly enough to form an ordered crystalline structure, with each oxygen atom collecting the four nearest of its neighbours and arranging them about itself in a rigid lattice. This structure results in a more open assembly, and hence a lower density, than the closely packed assembly of molecules in the liquid phase. For this reason, water is one of the few substances that is actually less dense in solid form than in the liquid state, dropping from 1,000 to 917 kilograms per cubic metre. It is the reason why ice floats rather than sinking, so that, during the winter, it develops as a sheet on the surface of lakes and rivers rather than sinking below the surface and accumulating from the bottom.

As water is warmed from the freezing point of 0° to 4° C (from 32° to 39° F), it contracts and becomes denser. This initial increase in density takes place because at 0° C a portion of the water consists of open-structured molecular arrangements similar to those of ice crystals. As the temperature increases, these structures break down and reduce their volume to that of the more closely packed polymeric structures of the liquid state. With further warming beyond 4° C, the water begins to expand in volume, along with the usual increase in intermolecular vibrations caused by thermal energy.

The ice crystal. At standard atmospheric pressure and at temperatures near 0° C, the ice crystal commonly takes the form of sheets or planes of oxygen atoms joined in a series of open hexagonal rings. The axis parallel to the hexagonal rings is termed the *c*-axis and coincides with the optical axis of the crystal structure. A view of the ice crystal along this axis is shown in Figure 2A.

The *c*-axis

When viewed perpendicular to the *c*-axis, as in Figure 2B, the planes appear slightly dimpled. The planes are stacked in a laminar structure that occasionally deforms by gliding, like a deck of cards. When this gliding deformation occurs, the bonds between the layers break, and the hydrogen atoms involved in those bonds must become attached to different oxygen atoms. In doing so, they migrate within the lattice, more rapidly at higher temperatures. Sometimes they do not reach the usual arrangement of two hydrogen atoms connected by covalent bonds to each oxygen atom, so that some oxygen atoms have only one or as many as three hydrogen bonds. Such oxygen atoms become the sites of electrical charge. The speed of crystal deformation depends on these readjustments, which in turn are sensitive to temperature. Thus the mechanical, thermal, and electrical properties of ice are interrelated.

PROPERTIES

Mechanical properties. Like any other crystalline solid, ice subject to stress undergoes elastic deformation, returning to its original shape when the stress ceases. However, if a shear stress or force is applied to a sample of ice for a long time, the sample will first deform elastically and will then continue to deform plastically, with a permanent alteration of shape. This plastic deformation, or creep, is of great importance to the study of glacier flow. It involves two processes: intracrystalline gliding, in which the layers within an ice crystal shear parallel to each other without destroying the continuity of the crystal lattice, and recrystallization, in which crystal boundaries change in size or shape depending on the orientation of the adjacent crystals and the stresses exerted on them. The motion of dislocations—that is, of defects or disorders in the crystal lattice—controls the speed of plastic deformation. Dislocations do not move under elastic deformation.

The strength of ice, which depends on many factors, is difficult to measure. If ice is stressed for a long time, it deforms by plastic flow and has no yield point (at which permanent deformation begins) or ultimate strength. For short-term experiments with conventional testing machines, typical strength values in bars are 38 for crushing, 14 for bending, 9 for tensile, and 7 for shear.

Thermal properties. The heat of fusion (heat absorbed on melting of a solid) of water is 334 kilojoules per kilogram. The specific heat of ice at the freezing point is 2.04 kilojoules per kilogram per degree Celsius. The thermal conductivity at this temperature is 2.24 watts per metre kelvin.

From *Proceedings of the Royal Society* (1958)

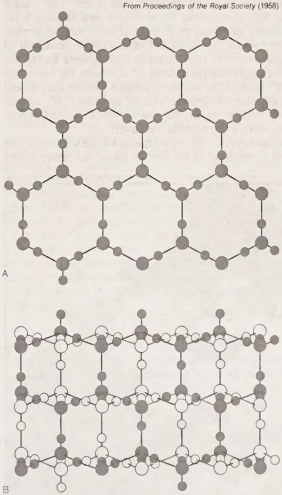


Figure 2: *Crystal structure of ice.* (A) View along the *c*-axis, showing the hexagonal formation adopted by the oxygen atoms (represented by large circles); (B) view perpendicular to the *c*-axis, showing the dimpled structure of the layered sheets of rings.

Another property of importance to the study of glaciers is the lowering of the melting point due to hydrostatic pressure: 0.0074° C per bar. Thus for a glacier 300 metres (984 feet) thick, everywhere at the melting temperature, the ice at the base is 0.25° C (0.45° F) colder than at the surface.

Optical properties. Pure ice is transparent, but air bubbles render it somewhat opaque. The absorption coefficient, or rate at which incident radiation decreases with depth, is about 0.1 cm^{-1} for snow and only 0.001 cm^{-1} or less for clear ice. Ice is weakly birefringent, or doubly refracting, which means that light is transmitted at different speeds in different crystallographic directions. Thin sections of snow or ice therefore can be conveniently studied under polarized light in much the same way that rocks are studied.

Albedo and radiative properties of ice

Electromagnetic properties. The albedo, or reflectivity (an albedo of 0 means that there is no reflectivity), to solar radiation ranges from 0.5 to 0.9 for snow, 0.3 to 0.65 for firn, and 0.15 to 0.35 for glacier ice. At the thermal infrared wavelengths, snow and ice are almost perfectly "black" (absorbent), and the albedo is less than 0.01. This means that snow and ice can either absorb or radiate long-wavelength radiation with high efficiency. At longer electromagnetic wavelengths (microwave and radio frequencies), dry snow and ice are relatively transparent, although the presence of even small amounts of liquid water greatly modifies this property. Radio echo sounding (radar) techniques are now used routinely to measure the thickness of dry polar glaciers, even where they are kilometres in thickness, but the slightest amount of liquid water distributed through the mass creates great difficulties with the technique. (G.D.A./M.F.M.)

Ice in lakes and rivers

Geographic extent. Much of the world experiences weather well below the freezing point, and in these regions ice forms annually in lakes and rivers. The nature of the ice formations may be as simple as a floating layer that gradually thickens, or it may be extremely complex, particularly when the water is fast-flowing. About half of the surface waters of the Northern Hemisphere freeze annually; the approximate extents of northern ice cover are shown in Figure 3. In warmer climates, waters may freeze only occasionally during periods of unusual cold, and in extremely cold areas of the world, such as Antarctica, lakes may have a permanent ice cover.

The seasonal cycle. In most regions where ice occurs, the formation is seasonal in nature: an initial ice cover forms some time after the average daily air temperature falls below the freezing point; the ice cover thickens through the winter period; and the ice melts and decays as temperatures warm in the spring. During the formation and thickening periods, energy flows out of the ice cover, and, during the decay period, energy flows into the ice cover. This flow of energy consists of two basic modes of energy exchange: (1) the radiation of long-wavelength and short-wavelength electromagnetic energy (*i.e.*, infrared and ultraviolet light) and (2) the transfer of heat energy associated with evaporation and condensation, with convection between the air and the surface, and (to a lesser extent) with precipitation falling on the surface. While radiation transfers are important, the dominant energy exchange in ice formation and decay is the heat transfer associated with evaporation and condensation and with turbulent convection—the latter being termed the sensible transfer. Since these transfers of heat are driven by the difference between air temperature and surface temperature, the extent and duration of ice covers more or less coincide with the extent and duration of average air temperatures below the freezing point (with a lag in the autumn due to the cooling of the water from its summer heating and a lag in the spring due to the melting of ice formed over the winter).

Energy transfer during ice formation and decay

As a general rule, small lakes freeze over earlier than rivers, and ice persists longer on lakes in the spring. Where there are sources of warm water—for example, in underground springs or in the thermal discharges of industrial power plants—this pattern may be disrupted, and water may be free of ice throughout the winter. In addition, in very deep lakes the thermal reserve built up during summer heating may be too large to allow cooling to the freezing point, or the action of wind over large fetches may prevent a stable ice cover from forming.

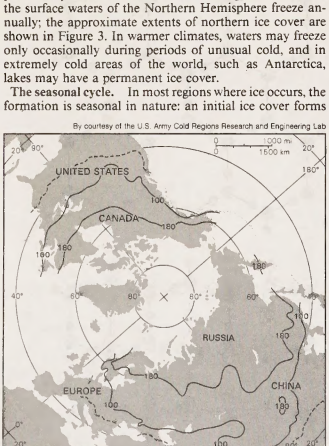


Figure 3: Distribution of ice cover in the Northern Hemisphere.

ICE IN LAKES

Ice formation. *Changes in temperature structure.* The setting for the development of ice cover in lakes is the annual evolution of the temperature structure of lake water. In most lakes during the summer, a layer of warm water of lower density lies above colder water below. In late summer, as air temperatures fall, this top layer begins to cool. After it has cooled and has reached the same density as the water below, the water column becomes isothermal (*i.e.*, there is a uniform temperature at all depths). With further cooling, the top water becomes even denser and plunges, mixing with the water below, so that the lake continues to be isothermal but at ever colder temperatures. This process continues until the temperature drops to that of the maximum density of water (about 4° C, or 39° F). Further cooling then results in expansion of the space between water molecules, so that the water becomes less dense. This change in density tends to create a new stratified thermal structure, this time with colder, lighter water on top of the warmer, denser water. If there is no mixing of the water by wind or currents, this top layer will cool to the freezing point (0° C, or 32° F). Once it is at the freezing point, further cooling will result in ice formation at the surface. This layer of ice will effectively block the exchange of energy between the cold air above and the warm water below; therefore, cooling will continue at the surface, but, instead of dropping the temperature of the water below, the heat losses will be manifested in the production of ice.

Changes in density of cooling water

The simple logic outlined above suggests that water at some depth in lakes during the winter will always be at 4° C, the temperature of maximum density, and indeed this is often the case in smaller lakes that are protected from the wind. The more usual scenario, however, is that wind mixing continues as the water column cools below 4° C, thereby overcoming the tendency toward density stratification. Between 4° and 0° C, for example, the density difference might be only 0.13 kilogram per cubic metre (3.5 ounces per cubic yard). Eventually some particular combination of cold air temperature, radiation loss, and low wind allows a first ice cover to form and thicken sufficiently to withstand wind forces that may break it up. As a result, even in fairly deep lakes the water temperature beneath the ice is usually somewhere below 4° C and

quite often closer to 0° C. The temperature at initial ice formation may vary from year to year depending on how much cooling has occurred before conditions are right for the first initial cover to form and stabilize. In some large lakes, such as Lake Erie in North America, wind effects are so great that a stable ice cover rarely forms over the entire lake, and the water is very near 0° C throughout the winter.

Nucleation of ice crystals. Before ice can form, water must supercool and ice crystals nucleate. Homogeneous nucleation (without the influence of foreign particles) occurs well below the freezing point, at temperatures that are not observed in water bodies. The temperature of heterogeneous nucleation (nucleation beginning at the surface of foreign particles) depends on the nature of the particles, but it is generally several degrees below the freezing point. Again, supercooling of this magnitude is not observed in most naturally occurring waters, although some researchers argue that a thin surface layer of water may achieve such supercooling under high rates of heat loss. Nucleation beginning on an ice particle, however, can take place upon only slight supercooling, and it is generally believed that ice particles originating from above the water surface are responsible for the initial onset of ice on the surface of a lake. Once ice is present, further formation is governed by the rate at which the crystal can grow. This can be very fast: on a cold, still night, when lake water has been cooled to its freezing point and then slightly supercooled on the surface, it is possible to see ice crystals propagating rapidly across the surface. Typically, this form of initial ice formation is such that the crystal *c*-axes are vertically oriented—in contrast to the usual horizontal orientation of the *c*-axis associated with later thickening. Under ideal conditions these first crystals may have dimensions of one metre or more. An ice cover composed of such crystals will appear black and very transparent.

Initial ice formation on lake surfaces

Effects of wind mixing. If the lake surface is exposed to wind, the initial ice crystals at the surface will be mixed by the agitating effects of wind on the water near the surface, and a layer of small crystals will be created. This layer will act to reduce the mixing, and a first ice cover will be formed consisting of many small crystals. Whether it is composed of large or small crystals, the ice cover, until it grows thick enough to withstand the effects of later winds, may form and dissipate and re-form repeatedly. On larger lakes where the wind prevents a stable ice cover from initially forming, large floes may be formed, and the ice cover may ultimately stabilize as these floes freeze together, sometimes forming large ridges and piles of ice. Ice ridges generally have an underwater draft several times their height above water. If they are moved about by the wind, they may scour the bottom in shallower regions. In some cases—particularly before a stable ice cover forms—wind mixing may be sufficient to entrain ice particles and supercooled water to considerable depths. Water intakes tens of metres deep have been blocked by ice during such events.

Ice growth. *Rates of growth.* Once an initial layer of ice has formed at the lake surface, further growth proceeds in proportion to the rate at which energy is transferred from the bottom surface of the ice layer to the air above. Because at standard atmospheric pressure the boundary between water and ice is at 0° C, the bottom surface is always at the freezing point. If there is no significant flow of heat to the ice from the water below, as is usually the case, all the heat loss through the ice cover will result in ice growth at the bottom. Heat loss through the ice takes place by conduction; designated φ in Figure 4, it is proportional to the thermal conductivity of the ice (k_i) and to the temperature difference between the bottom and the top surface of the ice ($T_m - T_s$), and it is inversely proportional to the thickness of the ice (h). Heat loss to the air above (also designated φ) occurs by a variety of processes, including radiation and convection, but it may be characterized approximately by a bulk transfer coefficient (H_{ia}) times the difference between the surface temperature of the ice and the air temperature ($T_s - T_a$). (In practice, the top surface of an ice layer is not at the air temperature but somewhere between the air temperature and the freezing point. The exact figures are rarely available, but fortunately the top surface temperature, T_s, is not needed for analysis.)

Heat loss through an ice layer

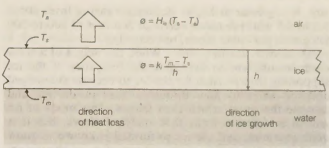


Figure 4: Heat flow through an ice cover (see text).

Assuming that the heat flow through the ice equals the heat flow from the surface of the ice to the air above, the following formula for the thickening of ice may be fashioned:

$$h = \left[\frac{2k}{\rho_i L}(T_m - T_a)t + \left(\frac{k}{H_{ia}}\right)^2\right]^{1/2} - \frac{k}{H_{ia}}. \qquad (1)$$

In this formula h is the thickness of the ice, T_a is the air temperature, T_m is the freezing point, k is the thermal conductivity of ice (2.24 watts per metre kelvin), ρ_i is the density of ice (916 kilograms per cubic metre), L is the latent heat of fusion (3.34×10^5 joules per kilogram), and t is the time since initial ice formation. The exact value of the bulk transfer coefficient (H_{ia}) depends on the various components of the energy budget, but it usually falls between 10 and 30 watts per square metre kelvin. Higher values are associated with windy conditions and lower values with still air conditions, but, with other information unavailable, a value of 20 watts per square metre kelvin fits data on ice growth quite well. Table 1 presents the results of calculations based on the above formula. The formula is particularly useful in predicting growth when the ice cover is thin. The first growth rate of the ice cover is proportional to the time since formation; as the ice thickens, however, the top surface temperature more closely approaches the air temperature, and growth proceeds proportional to the square root of time.

If there is a snow layer on top of the ice, it will offer a resistance to the flow of heat from the bottom of the ice surface to the air above. In this case, the incremental thickening rate (that is, the incremental thickening [dh] in an incremental time period [dt]) may be predicted by the following formula:

$$\frac{dh}{dt} = \frac{1}{\rho L}\left(\frac{T_m - T_a}{h_s/k_s + h_i/k_i + 1/H_{ia}}\right), \qquad (2)$$

where h_i is now the ice thickness with thermal conductivity k_i, and h_s is the snow thickness with thermal conductivity k_s. The thermal conductivity of snow depends on its den-

Table 1: Time Required for Lake Ice Formation

thickness of ice cover		air temperature			
cm	in.	−5° C (23° F)	−10° C (14° F)	−15° C (5° F)	−20° C (−4° F)
1	0.4	8.9 hours	4.4 hours	2.2 hours	1.5 hours
2	0.8	18.5 hours	9.3 hours	4.6 hours	3.1 hours
5	2	2.2 days	1.1 days	13.0 hours	8.7 hours
10	4	5.1 days	2.6 days	1.3 days	20.5 hours
20	8	13.4 days	6.7 days	3.4 days	2.2 days
50	20	57 days	29 days	14 days	9.5 days

sity. It is greater at higher densities, ranging from about 0.1 to 0.5 watt per metre kelvin at densities of 200 to 500 kilograms per cubic metre, respectively.

Snow ice

Variations in ice structure. When the weight of a snow cover is sufficient to overcome the buoyancy of the ice supporting it, it is usual for the ice to become submerged and for water to flow through cracks in the ice and saturate the snow, which then freezes. This mode of ice growth is different from that analyzed above, but it is quite common, and the ice so formed is known as snow ice. At typical snow densities, a layer of snow about one-half the thickness of the supporting ice will result in the formation of snow ice layers.

As the ice thickens, there is a tendency for crystals with a horizontal *c*-axis orientation to wedge out adjacent crystals with a vertical *c*-axis orientation and so become larger in diameter with depth. The resulting structure is one of adjacent columns of single crystals and is termed columnar ice. When a very thin section of the ice is cut and examined with light through crossed polaroid sheets, the crystal structure is clearly seen (see Figure 5).



Figure 5: Thin section of lake ice, showing the crystal structure of snow ice above the columnar crystal structure of the thickening ice cover.

Ice decay. *Thinning and rotting.* In the spring, when average daily air temperatures rise above the freezing point, ice begins to decay. Two processes are active during this period: a dimensional thinning and a deterioration of the ice crystal grains at their boundaries. Thinning of the ice layer is caused by heat transfer and by melting at the top or bottom surface (or both). Deterioration, sometimes called rotting or candling because of the similarity of deteriorating ice crystals to an assembly of closely packed candles, is caused by the absorption of solar radiation. When energy from the Sun warms the ice, melting begins at the grain boundaries because the melting point there is depressed by the presence of impurities that have been concentrated between crystal grains during the freezing process. Rotting may begin at the bottom or at the top, depending on the particular thermal conditions, but eventually the ice rots throughout its thickness. This greatly reduces the strength of the ice, so that rotten ice will support only a fraction of the load that solid, unrotted ice will support. Thinning and deterioration may occur simultaneously or independently of each other, so that sometimes ice thins without internal deterioration, and sometimes it deteriorates internally with little or no overall thinning. However, both processes usually occur before the ice cover finally breaks up.

Deteriorating ice has a gray, blotchy appearance and looks rotten. Because rotting takes place only by absorption of solar radiation, it progresses only during daylight hours. In addition, the presence of snow or snow ice, which either reflects most solar radiation or absorbs it rapidly in a thin layer, acts to prevent rotting of the ice below until the snow has been completely melted.

Melting. Melting of lake ice usually occurs first near the shorelines or near the mouths of streams. At these points of contact with inflowing warm water, the ice melts faster than it does at central lake locations, where most melting is caused by the transfer of heat from the atmosphere. Estimates of the rate at which thinning of the main ice cover occurs are usually based on a temperature index method in which a coefficient is applied to the air temperature above freezing.

Water temperature beneath the ice usually reaches its coldest at the time of freeze-up and then gradually warms throughout the winter. The warming is caused by the absorption of some solar radiation that has penetrated the ice cover, by the release of heat that has been stored in bottom sediments during the previous summer, and by warm water inflows. In deep lakes such warming is slight, while in shallow lakes it may amount to several degrees. After snow on the ice has melted in the spring, more solar radiation penetrates the ice cover, so that significant warming may occur. The mixing of warmed water with deteriorated ice is responsible for the very rapid clearing of lake ice at the end of the melt season. On most lakes, the timing of the final clearing of ice is remarkably uniform from year to year, usually varying by less than a week from the long-term average date of clearing.

Sources of heat during spring thaw

Geographic distribution. *Freeze-up.* The first appearance of lake ice follows by about one month the date at which the long-term average daily air temperature first falls below freezing. Ice appears first in smaller shallow lakes, often forming and melting several times in response to the diurnal variations in air temperature, and finally forms completely as air temperatures remain below the freezing point. Larger lakes freeze over somewhat later because of the longer time required to cool the water. In North America the Canadian-U.S. border roughly coincides with a first freeze-up date of December 1. North of the border freeze-up occurs earlier, as early as October 1 at Great Bear Lake in Canada's Northwest Territories. To the south the year-to-year patterns of freeze-up are ever more erratic until, at latitudes lower than about 45° N, freeze-up may not occur in some years.

In Europe the freeze-up pattern is similar with respect to air temperatures, but the latitudinal pattern shows more variation because much of western Europe is affected by the warming influence of the Gulf Stream. In Central Asia the latitudinal variation is more regular, with first freeze-up occurring about mid-January at 45° N and about October 1 at 72° N. Exceptions to these patterns occur where there are variations in local climate and elevation.

Clearing. Because of the time required to melt ice that has thickened over the winter, the clearing of lake ice occurs some time after average daily air temperatures rise above freezing. Typically the lag is on the order of one month at latitude 50° N and about six weeks at 70° N. This pattern results in average clearing dates in mid-April at the U.S.-Canadian border and in June and July in the northern reaches of Canada.

ICE IN RIVERS

Formation and growth. *Ice particles.* The formation of ice in rivers is more complex than in lakes, largely because of the effects of water velocity and turbulence. As in lakes, the surface temperature drops in response to cooling by the air above. Unlike lakes, however, the turbulent mixing in rivers causes the entire water depth to cool uniformly even after its temperature has fallen below the temperature of maximum density (4° C, or 39° F). The general pattern is one in which the water temperature fairly closely follows the average daily air temperature but with diurnal variations smaller than the daily excursions of air temper-

ature. Once the water temperature drops to the freezing point and further cooling occurs, the water temperature will actually fall below freezing—a phenomenon known as supercooling. Typically the maximum supercooling that is observed is only a few hundredths of a degree Celsius. At this point the introduction of ice particles from the air causes further nucleation of ice in the flow. This freezing action releases the latent heat of fusion, so that the temperature of the water returns toward the freezing point. Ice production is then in balance with the rate of cooling occurring at the surface.

Frazil ice

The particles of ice in the flow are termed frazil ice. Frazil is almost always the first ice formation in rivers. The particles are typically about 1 millimetre (0.04 inch) or smaller in size and usually in the shape of thin disks. Frazil appears in several types of initial ice formation: thin, sheetlike formations (at very low current velocities); particles that appear to flocculate into larger masses and exhibit a slushlike appearance on the water surface; irregularly shaped "pans" of frazil masses that, while appearing to be shallow, are actually of some depth; and (at high current velocities) a dispersed mixture or slurry of ice particles in the flow.

The supercooling of river water, while amounting to only a few hundredths of a degree Celsius or even less, provides the context for the particles to stick to one another, since under such conditions ice particles are inherently unstable and actively grow into the supercooled water. When they touch one another or some other surface that is cooled below the freezing point, they adhere by freezing. This behaviour causes serious problems at water intakes, where ice particles may adhere and then build up large accumulations that act to block the intake. In rivers and streams, frazil particles also may adhere to the bottom and successively build up a loose, porous layer known as anchor ice. Conversely, if the water temperature then rises above the freezing point, the particles will become neutral and will not stick to one another, so that the flow will be merely one of solid particles in the flowing water. The slightly above-freezing water may also release the bond between anchor ice and the bottom: it is not unusual for anchor ice to form on the bottom of shallow streams at night, when the cooling is great, only to be released the following day under the warming influence of air temperature and solar radiation.

Alberta Government Photograph

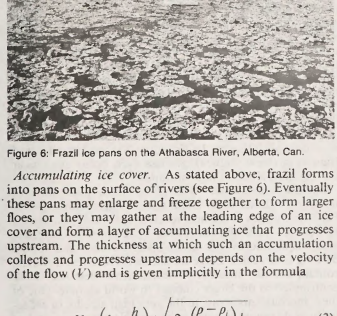


Figure 6: Frazil ice pans on the Athabasca River, Alberta, Can.

Accumulating ice cover. As stated above, frazil forms into pans on the surface of rivers (see Figure 6). Eventually these pans may enlarge and freeze together to form larger floes, or they may gather at the leading edge of an ice cover and form a layer of accumulating ice that progresses upstream. The thickness at which such an accumulation collects and progresses upstream depends on the velocity of the flow (V) and is given implicitly in the formula

$$V = \left(1 - \frac{h}{H}\right)\sqrt{2g\left(\frac{\rho - \rho_i}{\rho}\right)h}. \qquad (3)$$

in which g is acceleration of gravity, ρ and ρ_i are the densities of water and ice, respectively, h is the thickness of the accumulating ice, and H is the depth of flow just upstream of the ice cover. As a practical matter, floes arriving at the upstream edge will submerge and pass on downstream if the mean velocity exceeds about 60 centimetres (24 inches) per second. At certain thicknesses the ice accumulation may not be able to resist the forces exerted by the water flow and by its own weight acting in the downstream direction, and it will thicken by a shoving process until it attains a thickness sufficient to withstand these forces. During very cold periods, freezing of the top layer will provide additional strength by distributing the forces to the shorelines, so that thinner ice covers actually may be better able to withstand the forces acting on them.

Storage of water upstream

As the ice cover accumulates and progresses upstream, it both adds resistance to the flow and displaces a certain volume of water. These two effects cause the depth of the river to be greater upstream, thus reducing the velocity and enabling further upstream progression to occur where previously the current velocity was too high to allow ice cover formation. This phenomenon is termed staging, by reference to its effect of increasing the water level, or "stage." In the process there is a storage of water in the increased depth of the flow upstream, and this somewhat reduces the delivery of water downstream. The breakup of ice in the spring has the opposite effect—that is, the stored water is released and may contribute to a surge of water downstream.

Growth of fixed ice cover. Once the first ice cover has formed and stabilized, further growth is the same as with lake ice: typically columnar crystals grow into the water below, forming a bottom surface that is very smooth. This thickening may be predicted using equation (1), presented above for calculating the thickness of lake ice. An exception to this pattern arises when slightly above-freezing water flows beneath the ice cover. When this occurs, the action of the moving water either causes the undersurface to melt or retards the thickening. Since the rate at which melting occurs is proportional to the velocity times the water temperature, the ice cover over areas of higher velocity may be much thinner than in areas of lower velocity. Unfortunately, areas of thinner ice are often not apparent from above and may be dangerous to those traversing it.

In some rivers the initial formation of fixed ice takes place along the shorelines, with the central regions open to the air. The shore ice then gradually widens from the shoreline, and either the central region forms as described above by accumulation of frazil or the two sides of shore ice join.

Hanging dams

Ice buildups. In larger, deeper rivers, frazil produced in upstream reaches may be carried downstream and be transported beneath the fixed ice cover, where it may deposit and form large accumulations that are called hanging dams. Such deposits may be of great depth and may actually block large portions of the river's flow. In smaller, shallower streams, similar ice formations may be combinations of shore ice, anchor ice deposits, small hanging-dam-like accumulations, and (over slower-flowing areas) sheet ice.

Ice in smaller streams shows more variation through the winter, since most of the water comes from groundwater inflows during periods between rain. Groundwater is warm and over time may melt the ice formed during very cold periods. At other times all the water in a small stream freezes; subsequent inflowing water then flows over the surface and freezes, forming large buildups of ice. These are known as icings, *Aufeis* (German), or *naleds* (Russian). Icings may become so thick that they completely block culverts and in some cases overflow onto adjacent roads.

Decay and ice jams. In late winter, as air temperatures rise above the freezing point, river ice begins to melt owing to heat transfer from above and to the action of the slightly warm water flowing beneath. As occurs in lake ice, river ice also may deteriorate and rot because of absorption of solar radiation. On the undersurface, the action of the turbulent flowing water causes a melt pattern in the form of a wavy relief, with the waves oriented crosswise to the current direction. Eventually, if the ice cover is

not subjected to a suddenly increased flow, it may melt in place with little jamming or significant rise in water level. More likely, however, the ice may be moved and form ice jams.

During the spring in very northern areas, and during periods of midwinter thaw in more temperate areas, additional runoff from snowmelt and rain increases the flow in the river. The increased flow raises the water level and may break ice loose from the banks. It also increases the forces exerted on the ice cover. If these forces exceed the strength of the ice, the cover will move and break up and be transported downstream. At some places the quantity of ice will exceed the transport capacity of the river, and an ice jam will form. The jam may then build to thicknesses great enough to raise the water level and cause flooding. Typically, jams form where the slope of the river changes from steeper to milder or where the moving ice meets an intact ice cover—as in a large pool or at the point of outflow into a lake.

Spring breakup jams are usually more destructive than freeze-up jams because of the larger quantities of ice present. Besides causing sudden flooding, the ice itself may collide with structures and cause damage, even to the point of taking out bridges. Sometimes a jam forms, water builds up above it, and the jam breaks loose and moves downstream only to form again. This process may repeat itself several times. In northerly flowing rivers such behaviour is typical, since the upstream ice is freed first and moves toward colder, more stable ice covers.

Ice modification. *Mechanical methods.* There are a variety of means of modifying ice in rivers. Icebreaking vessels are used to clear paths for other vessels and occasionally to assist in relieving jams on large rivers. Icebreakers are used extensively in northern Europe and to some extent on the Great Lakes and St. Lawrence River of North America. Dusting the ice cover with a dark material such as coal dust or sand can increase the absorption of solar radiation and thus create areas of weakness that aid in an orderly breakup. Dusting has limited effectiveness, however, if a later snowfall covers the dust layer. Trenching of the ice cover with a ditching or similar machine has been practiced to create a weak zone in areas that are historically prone to jamming. Once ice jams have formed, they are sometimes blasted with explosives; however, if there is no current to transport the ice away after blasting, such measures are usually of little effect.

Difficulty of ice control

Ice-retention structures such as floating ice booms are used to hold ice in place and prevent it from moving downstream, where it might cause problems. There have been some attempts to control water releases from dammed reservoirs so as to induce breakup in an orderly manner, but these measures are limited to a narrow range of conditions. Air bubbler systems and flow developers (submerged motor-driven propellers) are used to melt small portions of the ice cover by taking advantage of any thermal reserve, relative to the freezing point, that may exist in the water. These are usually more successful in lakes or enclosed areas than in rivers, since the water temperature in rivers is rarely much above the freezing point.

Thermal methods. Wastewater from the cooling of power plants, both fossil-fueled and nuclear, has sometimes been suggested as a source of energy for melting ice downstream of the release points. This method may be advantageous in small areas, but the power requirements for melting extended reaches of ice are immense. Discharges from smaller sources, such as sewage treatment plants, are generally too small to have more than a very localized effect. On the other hand, the water held in reservoirs is often somewhat warmer than freezing, and it can be released in quantities sufficient to result in extended open water downstream—the precise distance depending on how much surface area is required to cool the water back to the freezing point by heat loss to the cold air above.

Geographic distribution. Dates of first freeze-up of rivers follow patterns similar to those of lakes, with a tendency for rivers to freeze over somewhat later than smaller lakes. The many factors that affect the freezing process of rivers make generalizations difficult, however. Slower poollike reaches may freeze over, while more rapidly flowing reaches may remain open well into the winter. Breakup is even more erratic, particularly in the more temperate zones where midwinter thaws may cause a breakup that is followed by another freeze-up and a later breakup as spring temperatures arrive. As a general rule, rivers break up in response to runoff from snowmelt or rain well before lakes clear of ice—although the first shoreline melting in lakes occurs at about the same times as river breakup. In north-flowing rivers, especially in central Russia and western Canada, breakup occurs first in upstream, southerly reaches and then progresses northward with the movement of the spring thaw. (G.D.A.)

Glaciers and ice sheets

A glacier may be defined as a large mass of perennial ice that originates on land by the recrystallization of snow or other forms of solid precipitation and that shows evidence of past or present flow. The definition is not precise, because exact limits for the terms large, perennial, and flow cannot be set. Except in size, a small snow patch that persists for more than one season is hydrologically indistinguishable from a true glacier. One international group has recommended that all persisting snow and ice masses larger than 0.1 square kilometre (about 0.04 square mile) be counted as glaciers.

GENERAL OBSERVATIONS

Main types of glaciers. Glaciers are classifiable in three main groups: (1) glaciers that extend in continuous sheets, moving outward in all directions, are called ice sheets if they are the size of Antarctica or Greenland and ice caps if they are smaller; (2) glaciers confined within a path that directs the ice movement are called mountain glaciers; and (3) glaciers that spread out on level ground or on the ocean at the foot of glaciated regions are called piedmont glaciers or ice shelves, respectively. Glaciers in the third group are not independent and are treated here in terms of their sources: ice shelves with ice sheets, piedmont glaciers with mountain glaciers. A complex of mountain glaciers burying much of a mountain range is called an ice field.

Distribution of glaciers. A most distinctive aspect of the last ice age, the Pleistocene Epoch (from 1,600,000 to 10,000 years ago), was the recurrent expansion and contraction of the world's ice cover. These glacial fluctuations influenced geological, climatological, and biological environments and affected the evolution and development of early humans. Pleistocene chronology is based primarily on ice-sheet fluctuations. Almost all of Canada, the northern third of the United States, much of Europe, all of Scandinavia, and large parts of northern Siberia were engulfed by ice during the major glacial stages. At times during the Pleistocene Epoch, glacial ice covered 30 percent of the world's land area; at other times the ice cover may have shrunk to less than its present extent. It may not be improper, then, to state that the world is still in an ice age. Because the term glacial generally implies ice-age events or Pleistocene time, in this discussion "glacier" is used as an adjective whenever reference is to ice of the present day.

Water content of the world's glaciers

Glacier ice today stores about three-fourths of all the fresh water in the world. Glacier ice covers about 11 percent of the world's land area and would cause a world sea-level rise of about 90 metres (300 feet) if all existing ice melted. Glaciers occur in all parts of the world and at almost all latitudes. In Ecuador, Kenya, Uganda, and Irian Jaya (New Guinea), glaciers even occur at or near the Equator, albeit at high altitudes.

Glaciers and climate. The cause of the fluctuation of the world's glacier cover is still not completely understood. Periodic changes in the heat received from the Sun, caused by fluctuations in the Earth's orbit, are known to correlate with major fluctuations of ice sheet advance and retreat on long time scales. Large ice sheets themselves, however, contain several "instability mechanisms" that may have contributed to the larger changes in world climate. One of these mechanisms is due to the very high albedo, or reflectivity of dry snow to solar radiation. No other material of widespread distribution on the Earth even approaches the

albedo of snow. Thus, as an ice sheet expands it causes an ever larger share of the Sun's radiation to be reflected back into space, less is absorbed on the Earth, and the world's climate becomes cooler. Another instability mechanism is implied by the fact that the thicker and more extensive an ice sheet is, the more snowfall it will receive in the form of orographic precipitation (precipitation resulting from the higher altitude of its surface and attendant lower temperature). A third instability mechanism has been suggested by studies of the West Antarctic Ice Sheet. Portions of an ice sheet may periodically surge (move rapidly) outward, perhaps because of the buildup of a thick layer of wet, deformable material under the ice. Although the ultimate causes of ice ages are not known with certainty, scientists agree that the world's ice cover and climate are in a state of delicate, interactive balance.

Only the largest ice masses directly influence global climate, but all ice sheets and glaciers respond to changes in local climate—particularly changes in air temperature or precipitation. The fluctuations of these glaciers in the past can be inferred by features they have left on the landscape. By studying these features, researchers can infer earlier climatic fluctuations.

FORMATION AND CHARACTERISTICS OF GLACIER ICE

Transformation of snow to ice. Glacier ice is an aggregate of irregularly shaped, interlocking single crystals that range in size from a few millimetres to several tens of centimetres. Many processes are involved in the transformation of snowpacks to glacier ice, and they proceed at a rate that depends on wetness and temperature. Snow crystals in the atmosphere are tiny hexagonal plates, needles, stars, or other intricate shapes. In a deposited snowpack these intricate shapes are usually unstable, and molecules tend to evaporate off the sharp (high curvature) points of crystals and be condensed into hollows in the ice grains. This causes a general rounding of the tiny ice grains so that they fit more closely together. In addition, the wind may break off the points of the intricate crystals and thus pack them more tightly. Thus, the density of the snowpack generally increases with time from an initial low value of 50–250 kilograms per cubic metre (3–15 pounds per cubic foot). The process of evaporation and condensation may continue: touching grains may develop necks of ice that connect them (sintering) and that grow at the expense of other parts of the ice grain, or individual small grains may rotate to fit more tightly together. These processes proceed more rapidly at temperatures near the melting point and more slowly at colder temperatures, but they all result in a net densification of the snowpack. On the other hand, if a strong temperature gradient is present, water molecules may migrate from grain to grain, producing an array of intricate crystal shapes (known as depth hoar) of lowered density. If liquid water is present, the rate of change is many times more rapid because of the melting of ice from grain extremities with refreezing elsewhere, the compacting force of surface tension, refreezing after pressure melting (regulation), and the freezing of water between grains.

Densification of the snowpack

This densification of the snow proceeds more slowly after reaching a density of 500–600 kilograms per cubic metre, and many of the processes mentioned above become less and less effective. Recrystallization under stress caused by the weight of the overlying snow becomes predominant, and grains change in size and shape in order to minimize the stress on them. This change usually means that large or favourably oriented grains grow at the expense of others. Stresses due to glacier flow may cause further recrystallization. These processes thus cause an increase in the density of the mass and in the size of the average grain.

When the density of the aggregate reaches about 830 to 840 kilograms per cubic metre, the air spaces between grains are sealed off, and the material becomes impermeable to fluids. With time and the application of stress, the density rises further by the compression of air bubbles (see Figure 7). Only rarely in mountain glaciers does the density exceed 900 kilograms per cubic metre, but at great depths in ice sheets the density may approach that of pure ice (917 kilograms per cubic metre at 0° C and atmospheric pressure).

Firn (névé)

Snow that has survived one melting season is called firn (or névé); its density usually is greater than 500 kilograms per cubic metre in temperate regions but can be as low as 300 kilograms per cubic metre in polar regions. The permeability change at a density of about 840 kilograms per cubic metre marks the transition from firn to glacier ice. The transformation may take only three or four years and less than 10 metres of burial in the warm and wet environment of Washington state in North America, but high on the plateau of Antaretica the same process takes several thousand years and burial to depths of about 150 metres.

Mass balance. Glaciers are nourished mainly by snowfall, and they waste away by melting and runoff or by the breaking off of icebergs (calving). In order for a glacier to remain at a constant size, there must be a balance between income (accumulation) and outgo (ablation). If this mass balance is positive (more gain than loss), the glacier will grow; if it is negative, the glacier will shrink.

Accumulation refers to all processes that contribute mass to a glacier. Snowfall is predominant, but additional contributions may be made by hoarfrost (direct condensation of ice from water vapour), rime (freezing of supercooled water droplets on striking a surface), hail, the freezing of rain or meltwater, or avalanching of snow from adjacent slopes. Ablation refers to all processes that remove mass from a glacier. In temperate regions, melting at the surface

By courtesy of the U.S. Army, photographs, A.J. Gow

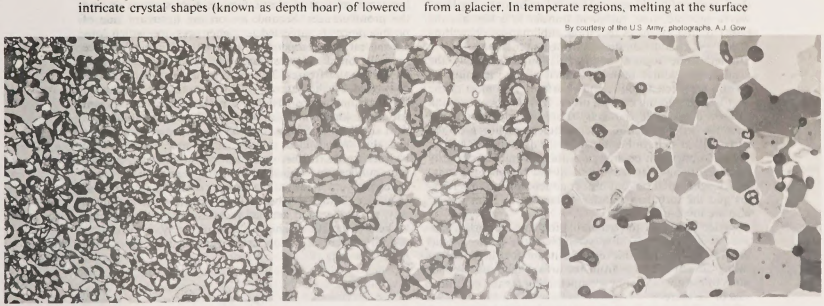


Figure 7: Thin-section photographs illustrating snow-to-ice transformation at depths of (left) 1 metre, (centre) 22 metres, and (right) 100 metres (magnified about 4.6×). At 1 metre and 22 metres the uniformly gray areas between grains of snow represent pore spaces. At 100 metres the snow has become ice and the bubbles become rounded and smaller because of compression of the entrapped gas. Photographs from Camp Century, Greenland, were taken between crossed polaroids to reveal crystal structure.

normally predominates. Melting at the base is usually very slight (1 centimetre [0.4 inch] per year or less). Calving is usually the most important process on large glaciers in polar regions and on some temperate glaciers as well. Evaporation and loss by ice avalanches are important in certain special environments; floating ice may lose mass by melting from below.

Because the processes of accumulation, ablation, and the transformation of snow to ice proceed so differently, depending on temperature and the presence or absence of liquid water, it is customary to classify glaciers in terms of their thermal condition. A polar glacier is defined as one that is below the freezing temperature throughout its mass for the entire year; a subpolar (or polythermal) glacier is mostly below the freezing temperature throughout its mass, except for surface melting in the summer; and a temperate glacier is at the melting temperature throughout its mass, but surface freezing occurs in winter. A polar or subpolar glacier may be frozen to its bed (cold-based), or it may be at the melting temperature at the bed (warm-based).

Another classification distinguishes the surface zones, or facies, on parts of a glacier. In the dry-snow zone no surface melting occurs, even in summer; in the percolation zone some surface melting may occur, but the meltwater refreezes at a shallow depth; in the soaked zone sufficient melting and refreezing take place to raise the whole winter snow layer to the melting temperature, permitting runoff; and in the superimposed-ice zone refrozen meltwater at the base of the snowpack (superimposed ice) forms a continuous layer that is exposed at the surface by the loss of overlying snow. These zones are all parts of the accumulation area, in which the mass balance is always positive. Below the superimposed-ice zone is the ablation zone, in which annual loss exceeds the gain by snowfall. The boundary between the accumulation and ablation zones is called the equilibrium line.

Accumulation and ablation zones

The value of the surface mass balance at any point on a glacier can be measured by means of stakes, snow pits, or cores. These values at points can then be averaged over the whole glacier for a whole year. The result is the net or annual mass balance. A positive value indicates growth, a negative value a decline.

Heat or energy balance. The mass balance and the temperature variations of a glacier are determined in part by the heat energy received from or lost to the external environment—an exchange that takes place almost entirely at the upper surface. Heat is received from short-wavelength solar radiation, long-wavelength radiation from clouds or water vapour, turbulent transfer from warm air, conduction upward from warmer lower layers, and the heat released by the condensation of dew or hoarfrost or by the freezing of liquid water. Heat is lost by outgoing long-wavelength radiation, turbulent transfer to colder air, the heat required for the evaporation, sublimation, or melting of ice, and conduction downward to lower layers.

In temperate regions, solar radiation is normally the greatest heat source (although much of the incoming radiation is reflected from a snow surface), and most of the heat loss goes to the melting of ice. It is incorrect to think of snow or ice melt as directly related to air temperature; it is the wind structure, the turbulent eddies near the surface, that determines most of the heat transfer from the atmosphere. In polar regions, heat is gained primarily from incoming solar radiation and lost by outgoing long-wavelength radiation, but heat conduction from lower layers and the turbulent transfer of heat to or from the air also are involved.

Glacier flow. In the accumulation area the mass balance is positive year after year. Here the glacier would become thicker and thicker were it not for the compensating flow of ice away from the area. This flow supplies mass to the ablation zone, compensating for the continual loss of ice there.

Glacier flow is a simple consequence of the weight and creep properties of ice. As stated above (see *Structure and properties of ice: Mechanical properties*), ice subjected to a shear stress over time will undergo creep, or plastic deformation. The rate of plastic deformation under constant

Rate of deformation

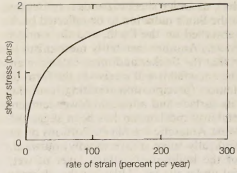


Figure 8: The rate of deformation, or shear strain, of ice at 0° C (32° F).

shear stress is initially high but tapers off to a steady value. If this steady value, the shear-strain rate, is plotted against the stress for many different values of applied stress, a curved graph will result (see Figure 8). The curve illustrates what is known as the flow law or constitutive law of ice: the rate of shear strain is approximately proportional to the cube of the shear stress. Often called the Glen flow law by glaciologists, this constitutive law is the basis for all analyses of the flow of ice sheets and glaciers.

As ice tends to build up in the accumulation area of a glacier, a surface slope toward the ablation zone is developed. This slope and the weight of the ice induce a shear stress throughout the mass. In a case with simple geometry, the shear stress can be given by the following formula:

$$\tau = \rho g h \sin \alpha \qquad (4)$$

where τ is the shear stress, ρ the ice density, h the ice thickness, and α the surface slope. Each element of ice deforms according to the magnitude of the shear stress, as determined by (4), at a rate determined by the Glen flow law, stated above. By adding up, or integrating, the shear deformation of each element throughout the glacier thickness, a velocity profile can be produced. It can be given numerical expression as:

$$u_1 = k_1 \sin^3 \alpha h^4 \qquad (5)$$

where u_1 is the surface velocity caused by internal deformation and k_1 a constant involving ice properties and geometry. In this simple case, velocity is approximately proportional to the fourth power of the depth (h^4). Therefore, if the thickness of a glacier is only slightly altered by changes in the net mass balance, there will be great changes in the rate of flow.

Glaciers that are at the melting temperature at the base may also slide on the bed. Two mechanisms operate to permit sliding over a rough bed. First, small protuberances on the bed cause stress concentrations in the ice, an increased amount of plastic flow, and ice streams around the protuberances. Second, ice on the upstream side of protuberances is subjected to higher pressure, which lowers the melting temperature and causes some of the ice to melt; on the downstream side the converse is true, and meltwater freezes. This process, termed regelation, is controlled by the rate at which heat can be conducted through the bumps. The first process is most efficient with large knobs, and the second process is most efficient with small bumps. Together these two processes produce bed slip. Water-filled cavities may form in the lee of bedrock knobs, further complicating the process. In addition, studies have shown that sliding varies as the basal water pressure or amount changes. Although the process of glacier sliding over bedrock is understood in a general way, none of several detailed theories has been confirmed by field observation. This problem is largely unsolved.

Sliding speed

A formula in common use for calculating the sliding speed is:

$$u_2 = \frac{k_2 \sin^2 \alpha}{(p_i - p_a)} \qquad (6)$$

where u_2 is the sliding speed at the base, p_i and p_a are the ice pressure and water pressure at the base of the ice, and k_2 is another constant involving a measure of the roughness of the bed. The total flow of a glacier can

thus be given by the sum of equations (5) and (6), u_i and u_2. The total sum would be an approximation, because the formulas ignore longitudinal changes in velocity and thickness and other complicating influences, but it has proved to be useful in analyzing situations ranging from small mountain glaciers to huge ice sheets.

Other studies have suggested that many glaciers and ice sheets do not slide on a rigid bed but "ride" on a deforming layer of water-charged sediment. This phenomenon is difficult to analyze because the sediment layer may thicken or thin, and thus its properties may change, depending on the history of deformation. In fact, the process may lead to an unsteady, almost chaotic, behaviour over time. Some ice streams in West Antarctica seem to have exhibited such unsteady behaviour.

Response of glaciers to climatic change. The relationship of glaciers and ice sheets to fluctuations in climate is sequential. The general climatic or meteorological environment determines the local mass and heat-exchange processes at the glacier surface, and these in turn determine the net mass balance of the glacier. Changes in the net mass balance produce a dynamic response—that is, changes in the rate of ice flow. The dynamic response causes an advance or retreat of the terminus, which may produce lasting evidence of the change in the glacier margin. If the local climate changes toward increased winter snowfall rates, the net mass balance becomes more positive, which is equivalent to an increase in ice thickness. The rate of glacier flow depends on thickness, so that a slight increase in thickness produces a larger increase in ice flow. This local increase in thickness and flow propagates down-glacier, taking some finite amount of time. When the change arrives at the terminus, it causes the margin of the glacier to extend farther downstream. The result is known as a glacier fluctuation—in this case an advance—and it incorporates the sum of all the changes that have taken place up-glacier during the time it took them to propagate to the terminus.

Glacier advance and retreat

The process, however, cannot be traced backward with assurance. A glacier advance can, perhaps, be related to a period of positive mass balances, but to ascertain the meteorological cause is difficult because either increased snowfall or decreased melting can produce a positive mass balance.

The dynamic response of glaciers to changes in mass balance can be calculated several ways. Although the complete, three-dimensional equations for glacier flow are difficult to solve for changes in time, the effect of a small change or perturbation in climate can be analyzed readily. Such an analysis involves the theory of kinematic waves, which are akin to small pulses in one-dimensional flow systems such as floods in rivers or automobiles on a crowded roadway. The length of time it takes the glacier to respond in its full length to a change in the surface mass balance is approximately given as the ratio of ice thickness to (negative) mass balance at the terminus. The time scale for mountain glaciers is typically on the order of 10 to 100 years—although for thick glaciers or those with low ablation rates it can be much longer. Ice sheets normally have time scales several orders of magnitude longer.

Glaciers and sea level. Sea level is currently rising at about 2 millimetres (0.08 inch) per year. Between 0.2 and 0.6 millimetre per year has been attributed to thermal expansion of ocean water, and most of the remainder is thought to be caused by the melting of glaciers and ice sheets on land. There is concern that the rate in sea-level rise may increase markedly in the future owing to global warming. Unfortunately, the state of the mass balance of the ice on the Earth is poorly known, so the exact contributions of the different ice masses to rising sea level is difficult to analyze. The mountain (small) glaciers of the world are thought to be contributing 0.2 to 0.4 millimetre per year to the rise. Yet the Greenland Ice Sheet is thought to be close to balance, the status of the Antarctic Ice Sheet is uncertain, and, although the floating ice shelves and glaciers may be in a state of negative balance, the melting of floating ice should not cause sea level to rise, and the grounded portions of the ice sheets seem to be growing. Thus, the cause of sea-level rise is an enigma.

With global warming, the melting of mountain glaciers will certainly increase, although this process is limited: the total volume of small glaciers is equivalent to only about 0.6 metre (2 feet) of sea-level rise. Melting of the marginal areas of the Greenland Ice Sheet will likely occur under global warming conditions, and this will be accompanied by the drawing down of the inland ice and increased calving of icebergs; yet these effects may be counterbalanced to some extent by increased snow precipitation on the inland ice. The Antarctic Ice Sheet, on the other hand, may actually serve as a buffer to rising sea level: increased melting of the marginal areas will probably be exceeded by increased snow accumulation due to the warmer air (which holds more moisture) and decreased sea ice (bringing moisture closer to the ice sheet). Modeling studies that predict sea-level rise up to the time of the doubling of greenhouse gas concentrations (*i.e.*, concentrations of atmospheric carbon dioxide, methane, nitrous oxide, and certain other gases) about the year 2050 suggest a modest rise of about 0.3 metre (1 foot).

Effects of global warming

THE GREAT ICE SHEETS

Two great ice masses, the Antarctic and Greenland ice sheets, stand out in the world today and may be similar in many respects to the large Pleistocene ice sheets. About 99 percent of the world's glacier ice is in these two ice masses, 91 percent in Antarctica alone.

Antarctic Ice Sheet. *Dimensions.* The bedrock of the continent of Antarctica is almost completely buried under ice (see Figure 9). Mountain ranges and isolated nunataks (a term derived from Greenland's Inuit language, used for individual mountains surrounded by ice) locally protrude through the ice. Extensive in area are the ice shelves, where the ice sheet extends beyond the land margin and spreads out to sea. The ice sheet, with its associated ice shelves, covers an area of 13,918,000 square kilometres (5,374,000 square miles); exposed rock areas total less than 200,000 square kilometres. The mean thickness of the ice is about 2,100 metres (6,900 feet), and the volume of ice more than 29 million cubic kilometres (7 million cubic miles). The land surface beneath the ice is below sea level in many places, but this surface is depressed because of the weight of the ice. If the ice sheet were melted, uplift of the land surface would eventually leave only a few deep troughs and basins below sea level—even though the sea level itself also would rise about 80 metres from the addition of such a large amount of water. Because of the thick ice cover, Antarctica has by far the highest mean altitude of the continents (2.2 kilometres [1.4 miles]); all other continents have mean altitudes less than 1 kilometre (0.6 mile).

Antarctic Peninsula. Antarctica can be divided into three main parts: the smallest and the mildest in climate is the Antarctic Peninsula, extending from latitude 63° S off the tip of South America to a juncture with the main body of West Antarctica at a latitude of about 74° S. The ice cover of the Antarctic Peninsula is a complex of ice caps, piedmont and mountain glaciers, and small ice shelves.

West Antarctica. The part of the main continent lying south of the Americas, between longitudes 45° W and 165° E, is characterized by irregular bedrock and ice-surface topography and numerous nunataks and deep troughs. Two large ice shelves occur in West Antarctica: the Ronne-Filchner Ice Shelf (often considered to be two separate ice shelves), south of the Weddell Sea, and the Ross Ice Shelf, south of the Ross Sea. Each has an area exceeding 400,000 square kilometres.

The great Antarctic ice shelves

East Antarctica. The huge ice mass of East Antarctica, about 10,200,000 square kilometres, is separated from West Antarctica by the Transantarctic Mountains. This major mountain range extends from the eastern margin of the Ross Ice Shelf almost to the Ronne-Filchner Ice Shelf. The bedrock of East Antarctica is approximately at sea level, but the ice surface locally exceeds 4,000 metres above sea level on the highest parts of the polar plateau.

Climatic conditions. At the South Pole the snow surface is 2,800 metres in altitude, and the mean annual temperature is about −50° C (−58° F), but at the Russian Vostok Station (78°27′ S, 106°52′ E), 3,500 metres above sea

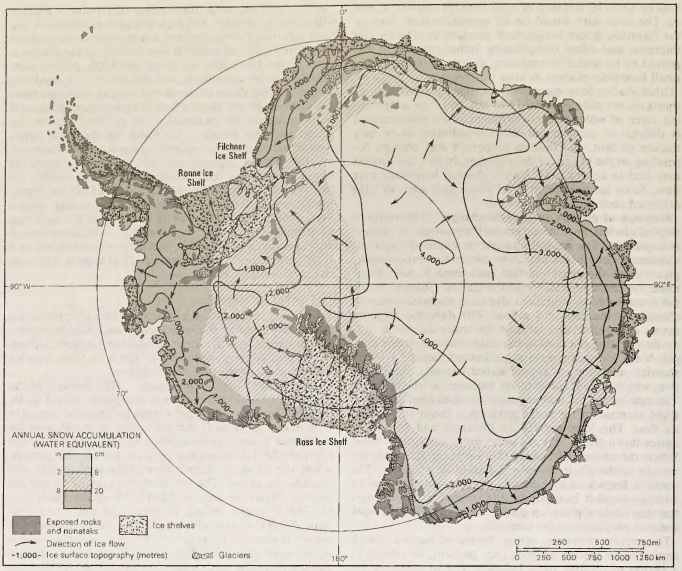


Figure 9: Nunataks, ice shelves, ice flow directions, and snow accumulation in Antarctica.
Adapted from M.B. Giovinetto, "Drainage Systems of Antarctica: Accumulation," *Antarctic Snow and Ice Studies*, Antarctic Research Series, vol. 2. American Geophysical Union (1964)

level, the mean annual temperature is −58° C (−73° F), and in August 1960 (the winter season) the temperature reached a low of −88.3° C (−127° F). The temperatures on the polar plateau of East Antarctica are by far the coldest on Earth; the climate of the Arctic is quite mild by comparison. Along the coast of East or West Antarctica, where the climate is milder, mean annual temperatures range from −20° to −9° C (−4° to 16° F), but temperatures exceed the melting point only for brief periods in summer, and then only slightly. Katabatic (drainage) winds, however, are very strong along the coast; the mean annual wind speed at Commonwealth Bay is 20 metres per second (45 miles per hour).

Greenland Ice Sheet. The Greenland Ice Sheet, though subcontinental in size, is huge compared with other glaciers in the world except that of Antarctica. Greenland is mostly covered by this single large ice sheet (1,730,000 square kilometres), while isolated glaciers and small ice caps totaling between 76,000 and 100,000 square kilometres occur around the periphery. The ice sheet is almost 2,400 kilometres long in a north-south direction, and its greatest width is 1,100 kilometres at a latitude of 77° N, near its northern margin. The mean altitude of the ice surface is 2,135 metres. The term Inland Ice, or, in Danish, Indlandsis, is often used for this ice sheet.

The bedrock surface is near sea level over most of the interior of Greenland, but mountains occur around the periphery. Thus, this ice sheet, in contrast to the Antarctic Ice Sheet, is confined along most of its margin. The ice surface reaches its greatest altitude on two north-south elongated domes, or ridges. The southern dome reaches almost 3,000 metres at latitudes 63°–65° N; the northern dome reaches about 3,290 metres at about latitude 72° N. The crests of both domes are displaced east of the centre line of Greenland.

The unconfined ice sheet does not reach the sea along a broad front anywhere in Greenland, so that no large ice shelves occur. The ice margin just reaches the sea, however, in a region of irregular topography in the area of Melville Bay southeast of Thule. Large outlet glaciers, which are restricted tongues of the ice sheet, move through bordering valleys around the periphery of Greenland to calve off into the ocean, producing the numerous icebergs that sometimes penetrate North Atlantic shipping lanes. The best known of these is the Jakobshavn Glacier, which, at its terminus, flows at speeds of 20 to 22 metres per day.

The climate of the Greenland Ice Sheet, though cold, is not as extreme as that of central Antarctica. The lowest mean annual temperatures, about −31° C (−24° F), occur on the north-central part of the north dome, and temperatures at the crest of the south dome are about −20° C (−4° F).

Mass balance of the ice sheets. *Accumulation.* The rate of precipitation on the Antarctic Ice Sheet is so low that it may be called a cold desert. Snow accumulation over much of the vast polar plateau is less than five centimetres (two inches) water equivalent per year. Only around the margin of the continent, where cyclonic storms penetrate frequently, does the accumulation rise to values of more than 30 centimetres. The mean for Antarctica is 15 centimetres or less. In Greenland values are higher: less than 15 centimetres in a comparatively small area of north-central Greenland, 30 centimetres along the crests of the domes, and more than 80 centimetres along the southeast and southwest margins; the mean annual snow accumulation is about 37 centimetres of water equivalent.

Annual snowfall

Snow accumulation occurs mainly as direct snowfall when cyclonic storms move inland. At high altitudes on the Greenland Ice Sheet and in central Antarctica, ice crystals form in the cold air during clear periods and slowly settle out as fine "diamond dust." Hoarfrost and rime deposition are generally minor items in the snow-

accumulation totals. It is almost impossible to measure the precipitation directly in these climates; precipitation gauges are almost useless for the measurement of blowing snow, and the snow is blown about almost constantly in some areas. The thickness and density of snow deposited on the ground equals precipitation plus hoarfrost and rime deposition, less evaporation, less snow blown away, and plus snow blown in from somewhere else. The last two phenomena are thought to cancel each other approximately—except in the coastal areas, where fierce drainage, or katabatic, winds move appreciable quantities of snow out to sea.

The snow surface may be smooth where soft powder snow is deposited with little wind, or very hard packed and rough when high winds occur during or after snowfall. Two features are prominent: snow dunes are depositional features resembling sand dunes in their several shapes; sastrugi are jagged erosional features (often cut into snow dunes) caused by strong prevailing winds that occur after snowfall. Sharp, rugged sastrugi, which can be one to two metres high, make travel by vehicle or on foot difficult. The annual snow layers exposed in the side of a snow pit can usually be distinguished by a low density layer (depth hoar) that forms by the burial of surface hoarfrost or by metamorphism of the snow deposited in the fall at a time when the temperature is changing rapidly.

Almost all of the Antarctic Ice Sheet lies within the dry-snow zone. The percolation, soaked, and superimposed ice zones occur only in a very narrow strip in a small area along the coast. In Greenland only the central part of the northern half of the ice sheet, or about 30 percent of the total area, is within the dry-snow zone, Almost half of the area of the Greenland ice sheet is considered to be in the percolation zone. In flat areas near the equilibrium line, especially in west-central Greenland, there are notorious snow swamps, or slush fields, in summer; some of this water runs off, but much of it refreezes. (For an explanation of a glacier's surface zones, see above *Formation and characteristics of glacier ice: Mass balance.*)

Ablation. The ice sheets lose material by several processes, including surface melting, evaporation, wind erosion (deflation), iceberg calving, and the melting of the bottom surfaces of floating ice shelves by warmer seawater.

Calving and melting of the Antarctic Ice Sheet

In Antarctica, calving of ice shelves and outlet glacier tongues clearly predominates among all the processes of ice loss, but calving is very episodic and cannot be measured accurately. The amount of surface melt and evaporation is small, amounting to about 22 centimetres of ice lost from a five-kilometre ring around half the continent. Wind erosion is difficult to evaluate but probably accounts for only a very small loss in the mass balance. The undersides of ice shelves near their outer margins are subject to melting by the ocean water. The rate of melting decreases inland, and at that point some freezing of seawater onto the base of the ice shelves must occur, but farther inland, near the grounding line, the tidal circulation of warm seawater may produce basal melting.

In Greenland, surface melt is more important, calving is less so, and undershelf melting is virtually nonexistent. Most of the calving is from the termini of a relatively few large, fast-moving outlet glaciers. In Greenland, vertical-walled melt pits in the ice are a well-known feature of the ice surface at the ablation zone. Ranging from a few millimetres to a metre in diameter, these pits are floored with a dark, silty material called cryoconite, once thought to be of cosmic origin but now known to be largely terrestrial dust. The vertical melting of the holes is due to the absorption of solar radiation by the dark silt, possibly augmented by biological activity.

Net mass balance. Because two great ice sheets contain 99 percent of the world's ice, it is important to know whether this ice is growing or shrinking under present climatic conditions. Although just such a determination was a major objective of the International Geophysical Year (1957–58) and more has been learned each year since, even the sign of the net mass balance has not yet been determined conclusively.

It appears that accumulation on the surface of the Antarctic Ice Sheet is approximately balanced by iceberg calving and basal melting from the ice shelves. Compilations published in 1990–93 suggest the following values, given in gigatons (billions of tons) per year (1 gigaton is equivalent to 1.1 cubic kilometres of water):

Accumulation	
Accumulation on grounded ice	+ 1,670 ± 330
Accumulation on ice shelves	+ 450 ± 90
Ablation	
Calving of ice shelves and glaciers	− 2,110 ± 700
Bottom melting, ice shelves	− 540 ± 180
Melting and runoff	− 50 ± 25
Net mass balance	− 580 ± 800

The net difference, however, is on the same order as the margin of error in estimating the various quantities. Furthermore, some authors have suggested that the values stated above for calving and ice-shelf melting are too high and that the discharge of ice to the sea, as measured by ice-flow studies, is clearly less than the accumulation. Thus, even the sign of the net balance is not well defined. It appears that the net balance of the grounded portion of the Antarctic Ice Sheet is positive, while that of the floating ice shelves is negative. Studies of fluctuations in the extent of floating ice have been inconclusive.

Mass balance of the Greenland Ice Sheet

The net mass balance of the Greenland Ice Sheet also appears to be close to zero, but here, too, the margin of error is too large for definite conclusions. The estimated balance is as follows, again in gigatons per year.

Accumulation	
Snow accumulation	+ 554 ± 83
Ablation	
Iceberg calving	− 311 ± 93
Melting and runoff	− 244 ± 73
Net mass balance	0 ± 145

Uncertainties in the quantities given above are due to the difficulty of analyzing the spatial and temporal distributions of accumulation, the relatively few annual measurements of iceberg calving, and a lack of knowledge of the amount of surface meltwater that refreezes in the cold snow and ice at depth. Numerous ice coring, leveling, and flow studies suggest that the Greenland Ice Sheet is increasing in thickness at a rate of up to 0.1 metre per year. One study of changes in thickness between 1978 and 1986, based on satellite measurement, indicates a much higher thickening rate, but this conclusion has not been independently verified. Many of the outlet glaciers and portions of the ice-sheet margin in the southwestern part of Greenland, where many observations have been made, have stopped the retreats that were observed from the 1950s through the 1970s; since the 1980s they have been stable or advancing.

Flow of the ice sheets. In general, the flow of the Antarctic and Greenland ice sheets is not directed radially outward to the sea. Instead, ice from central high points tends to converge into discrete drainage basins and then concentrate into rapidly flowing ice streams. (Such so-called streams are currents of ice that move several times faster than the ice on either side of them.) The ice of much of East Antarctica has a rather simple shape with several subtle high points or domes. Greenland resembles an elongated dome, or ridge, with two summits. West Antarctica is a complex of converging and diverging flow because of the jumble of ridges and troughs in the subglacial bedrock and the convergence of ice streams.

Flow rates

Flow rates in the interior of an ice sheet are very low, being measured in centimetres or metres per year, because the surface slope is minuscule and the ice is very cold. As the ice moves outward, the rate of flow increases to a few tens of metres per year, and this rate of flow increases still further, up to one kilometre per year, as the flow is channeled into outlet glaciers or ice streams. Ice shelves continue the flow and even cause it to increase, because ice spreads out in ever thinner layers. At the edge of the Ross Ice Shelf, ice is moving out about 900 metres per year toward the ocean.

This simple picture of ice flow is made more complicated by the dependence of the flow law of ice on temperature. Because a temperature increase of about 15° C (27° F)

causes a 10-fold increase in the deformation rate of ice, the temperature distribution of an ice sheet partly determines its flow structure. The cold ice of the central part of an ice sheet is carried down into warmer zones. This shift modifies the static temperature distribution, and the shear deformation is concentrated in a thin zone of warmer ice at the base. The forward velocity may be almost uniform throughout the depth to within a few tens or hundreds of metres from the bedrock.

Another important effect on ice flow is the heat produced by friction, caused by the sliding of the ice on bedrock or by internal shearing within the basal ice. If a portion of the ice sheet deforms more rapidly than its surroundings, the slight amount of extra heat production raises the temperature of this portion, causing it to deform even more readily. This increased deformation may explain the phenomena of ice streams. Ice streams are very effective in moving ice from large drainage areas of Antarctica and Greenland out to ice shelves or to the sea. It is known that at least one Antarctic ice stream moves rapidly on a layer of water-charged deforming sediment; a nearby ice stream appears to have ceased rapid movement in the past several hundred years, perhaps owing to loss of its sediment layer.

Information from deep cores. Most of the Antarctic Ice Sheet and much of the Greenland Ice Sheet are below freezing throughout. Continuous cores, taken in some cases to the bedrock below, allow the sampling of an ice sheet through its entire history of accumulation. Records obtained from these cores represent exciting new developments in paleoclimatology and paleoenvironmental studies. Because there is no melting, the layered structure of the ice preserves a continuous record of snow accumulation and chemistry, air temperature and chemistry, and fallout from volcanic, terrestrial, marine, cosmic, and man-made sources. Actual samples of ancient atmospheres are trapped in air bubbles within the ice. This record extends back more than 300,000 years.

Inspecting annual layers

Near the surface it is possible to pick out annual layers by visual inspection. In some locations, such as the Greenland Ice core Project/Greenland Ice Sheet Project 2 (GRIP/GISP2) sites at the summit of Greenland, these annual layers can be traced back more than 40,000 years, much like counting tree rings. The result is a remarkably high-resolution record of climatic change. When individual layers are not readily visible, seasonal changes in dust,

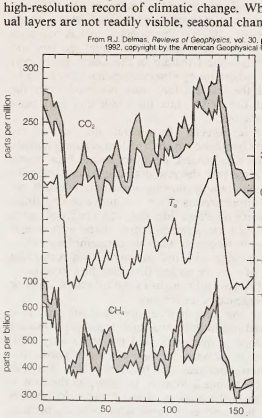


Figure 10: Parallel changes in the concentration of carbon dioxide (CO_2) and methane (CH_4) and in the temperature of the air over tens of thousands of years. Values are derived from measurements of oxygen and hydrogen isotope levels in air samples from a deep ice core taken at Vostok Station, East Antarctica.

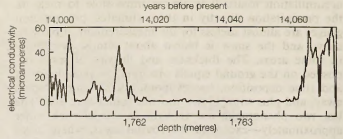


Figure 11: Fluctuations in electrical conductivity in an ice core taken by the Greenland Ice Sheet Project 2 in central Greenland. The changes in conductivity, caused by changes in atmospheric acidity brought about by windblown dust, indicate a rapid climatic transition between a warm period and a cold, glacial period.
From K.C. Taylor *et al.*, *Nature*, vol. 361, 1993

marine salts, and isotopes can be used to infer annual chronologies. Precise dating of recent layers can be accomplished by locating radioactive fallout from known nuclear detonations or traces of volcanic eruptions of known date. Other techniques must be used to reconstruct a chronology from some very deep cores. One method involves a theoretical analysis of the flow. If the vertical profile of ice flow is known, and if it can be assumed that the rate of accumulation has been approximately constant through time, then an expression for the age of the ice as a function of depth can be developed.

A very useful technique for tracing past temperatures involves the measurement of oxygen isotopes—namely, the ratio of oxygen-18 to oxygen-16. Oxygen-16 is the dominant isotope, making up more than 99 percent of all natural oxygen; oxygen-18 makes up 0.2 percent. However, the exact concentration of oxygen-18 in precipitation, particularly at high latitudes, depends on the temperature. Winter snow has a smaller oxygen-18–oxygen-16 ratio than does summer snow. A similar isotopic method for inferring precipitation temperature is based on measuring the ratio of deuterium (hydrogen-2) to normal hydrogen (hydrogen-1). The relation between these oxygen and hydrogen isotopic ratios, termed the deuterium excess, is useful for inferring conditions at the time of precipitation. The temperature scale derived from isotopic measurements can be calibrated by the observable temperature-depth record near the surface of ice sheets.

Results of ice core measurements are greatly extending the knowledge of past climates. For instance, air samples taken from ice cores show an increase in methane, carbon dioxide, and other "greenhouse gas" concentrations with the rise of industrialization and human population. On a longer time scale, the concentration of carbon dioxide in the atmosphere can be shown to be related to atmospheric temperature (as indicated by oxygen and hydrogen isotopes)—thus confirming the global-warming greenhouse effect, by which heat in the form of long-wave infrared radiation is trapped by atmospheric carbon dioxide and reflected back to the Earth's surface.

Evidence of climatic change

Perhaps most exciting are recent ice core results that show surprisingly rapid fluctuations in climate, especially during the last glacial period (160,000 to 10,000 years ago) and probably in the interglacial period that preceded it. Detectable variations in the dustiness of the atmosphere (a function of wind and atmospheric circulation), temperature, precipitation amounts, and other variables show that, during this time period, the climate frequently alternated between full-glacial and nonglacial conditions in less than a decade. Some of these changes seem to have occurred as "flickerings," in which the temperature jumped 5° to 7° C (9° to 13° F), remained in that state for a few years, jumped back, and repeated the process several times before settling into the new state for a long time—perhaps 1,000 years. These findings have profound and unsettling implications for the understanding of the coupled ocean-atmosphere climate system.

MOUNTAIN GLACIERS

In this discussion the term mountain glaciers includes all perennial ice masses other than the Antarctic and Greenland ice sheets. These ice masses are not necessarily asso-

ciated with mountains. Sometimes the term small glaciers is used, but only in a relative sense: a glacier 10,000 square kilometres (4,000 square miles) in surface area would not be called "small" in many parts of the world.

Classification of mountain glaciers. Mountain glaciers are generally confined to a more or less marked path directing their movement. The shape of the channel and the degree to which the glacier fills it determine the type of glacier. Valley glaciers are a classic type; they flow at least in part down a valley and are longer than they are wide. Cirque glaciers, short and wide, are confined to cirques, or amphitheatres, cut in the mountain landscape. Other types include transection glaciers or ice fields, which fill systems of valleys, and glaciers in special situations, such as summit glaciers, hanging glaciers, ice aprons, crater glaciers, and regenerated or reconstituted glaciers. Glaciers that spread out at the foot of mountain ranges are called piedmont glaciers. Outlet glaciers are valley glaciers that originate in ice sheets, ice caps, and ice fields. Because of the complex shapes of mountain landscapes and the resulting variety of situations in which glaciers can develop, it is difficult to draw clear distinctions among the various types of glaciers.

Mountain glaciers also are classified as polar, subpolar, or temperate and their surfaces by the occurrence of dry-snow, percolation, saturation, and superimposed-ice zones, as for ice sheets.

Surface features. The snow surface of the accumulation area of a mountain glacier displays the same snow dune and sastrugi features found on ice sheets, especially in winter, but normally these features are neither as large nor as well developed. Where appreciable melting of the snow occurs, several additional features may be produced. During periods of clear, sunny weather, sun cups (cup-shaped hollows usually between 5 and 50 centimetres [2 and 20 inches] in depth) may develop. On very high-altitude, low-latitude snow and firn fields these may grow into spectacular narrow blades of ice, up to several metres high, called nieves penitentes. Rain falling on the snow surface (or very high rates of melt) may cause a network of meltwater runnels (shallow grooves trending downslope) to develop.

Other features are characteristic of the ablation zone. Below icefalls (steep reaches of a valley glacier), several types of curved bands can be seen. The surface of the glacier may rise and fall in a periodic manner, with the spacing between wave crests approximately equal to the amount of ice flow in a year. Called wave ogives (pointed arches), these arcs result from the great stretching of the ice in the rapidly flowing icefall. The ice that moves through the icefall in summer has more of its surface exposed to melting and is greatly reduced in volume compared with the ice moving through in winter. Dirtband ogives also may occur below icefalls; these are caused by seasonal differences in the amount of dust or by snow trapped in the icefall. In plan view, the ogives are invariably distorted into arcs or curves convex downglacier; hence the name ogive.

Ogives and ice stratification

The ice of the ablation zone normally shows a distinctive layered structure. This can be relict stratification developed by the alternation of dense and light or of clean and dirty snow accumulations from higher on the glacier. This stratification is later subdued by recrystallization accompanying plastic flow. A new layering called foliation is developed by the flow. Foliation is expressed by alternating layers of clear and bubbly or coarse-grained and fine-grained ice. Although the origin of this structure is not fully understood, it is analogous to the process that produces foliated structures in metamorphic rocks.

The ice crystals in strongly deformed, foliated ice invariably have a preferred orientation, relative to the stress directions. In some situations, more often in polar than in temperate ice, the hexagonal axes are aligned perpendicularly to the plane of foliation. This alignment places the crystal glide planes parallel to the planes of (presumed) greatest shearing. In many other locations the hexagonal crystal axes are preferentially aligned in four different directions, none perpendicular to the foliation. This enigmatic pattern has resisted explanation so far.

Crevasses are common to both the accumulation and ablation zones of mountain glaciers, as well as of ice sheets. Transverse crevasses, perpendicular to the flow direction along the centre line of valley glaciers, are caused by extending flow. Splaying crevasses, parallel to the flow in midchannel, are caused by a transverse expansion of the flow. The drag of the valley walls produces marginal crevasses, which intersect the margin at 45°. Transverse and splaying crevasses curve around to become marginal crevasses near the edge of a valley glacier. Splaying and transverse crevasses may occur together, chopping the glacier surface into discrete blocks or towers, called seracs.

Crevasses deepen until the rate of surface stretching is counterbalanced by the rate of plastic flow tending to close the crevasses at depth. Thus, crevasse depths are a function of the rate of stretching and the temperature of the ice. Crevasses deeper than 50 metres (160 feet) are rare in temperate mountains, but crevasses to 100 metres or more in depth may occur in polar regions. Often the crevasses are concealed by a snow bridge, built by accumulations of windblown snow.

Accumulation and ablation on mountain glaciers

Mass balance of mountain glaciers. The rate of accumulation and ablation on mountain glaciers depends on latitude, altitude, and distance downwind from sources of abundant moisture, such as the oceans. The glaciers along the coasts of Washington, British Columbia, southeastern Alaska, South Island of New Zealand, Iceland, and southwestern Norway receive prodigious snowfall. Snow accumulation of three to five metres of water equivalent in a single season is not uncommon. With this large income, glaciers can exist at low altitudes in spite of very high melt rates. The rate of snowfall increases with increasing altitude; thus, the gradient of net mass balance with altitude is steep. This gradient also expresses the rate of transfer of mass by glacier flow from high to low altitudes and is called the activity index.

Typical of the temperate, maritime glaciers is South Cascade Glacier, in western Washington (see Figure 12). Its activity index is high, normally about 17 millimetres per metre (0.2 inch per foot); the yearly snow accumulation averages about 3.1 metres of water-equivalent; and the equilibrium line is at the relatively low altitude of 1,900 metres. This glacier contains only ablation and saturation zones; the winter chill is so slight that no superimposed ice is formed.

In the maritime environment of southeastern Alaska are many very large glaciers; Bering and Seward-Malaspina glaciers (piedmont glaciers) cover 5,800 and 5,200 square kilometres in area, respectively. Equilibrium lines are lower than those in Washington state, but the rates of accumulation and ablation and the activity indices are about the same. Because these mountains are high, and some glaciers extend over a great range of altitude, all surface zones except the dry-snow zone are represented.

In more continental (inland) environments, the rate of snowfall is much less, and the summer climate is generally warmer. Thus, glaciers can exist only at high altitudes. High winds may concentrate the meagre snowfall in deep, protected basins, however, allowing glaciers to form even in areas of low precipitation and high melt rates. Glaciers formed almost entirely of drift snow occur at high altitudes in Colorado and in the polar Ural Mountains and are often referred to as Ural-type glaciers. Superimposed ice and soaked zones are found in the accumulation area; in higher areas the percolation zone is found, and in some local extreme areas the dry-snow zone occurs. Because of the decrease in melt rates, continental glaciers in high latitudes occur at lower altitudes and have lower accumulation totals and activity indices. McCall Glacier, in the northwestern part of the Brooks Range in Alaska, has the lowest activity index (two millimetres per metre) measured in western North America. Glaciers in intermediate climates have intermediate equilibrium-line altitudes, accumulation or ablation totals, and activity indices.

Flow of mountain glaciers. Ice flow in valley glaciers has been studied extensively. The first measurements date from the mid-18th century, and the first theoretical analyses date from the middle of the 19th century. These glaciers generally flow at rates of 0.1 to 2 metres per day, faster at the surface than at depth, faster in midchannel than along the margins, and usually fastest at or just below



Figure 12: South Cascade Glacier, Washington state, U.S.; a typical small valley glacier.
By courtesy of the U.S. Geological Survey, photograph, Austin Post

the equilibrium line. Cold, polar glaciers flow relatively slowly, because the constitutive law of ice is sensitive to temperature and because they generally are frozen to their beds. In some high-latitude areas, such as the Svalbard archipelago north of Norway, polythermal glaciers are common; these consist of subfreezing ice overlying temperate ice, and, because they are warm-based, they actively slide on their beds.

The fastest glaciers

The fastest glaciers (other than those in the act of surging) are thick, temperate glaciers in which high subglacial water pressures produce high rates of sliding. Normal temperate glaciers ending on land generally have subglacial water pressures in the range of 50 to 80 percent of the ice pressure, but glaciers that end in the sea may have subglacial water pressures almost equal to the ice pressure—that is, they almost float. The lower reach of Columbia Glacier in southern Alaska, for instance, flows between 5 and 25 metres per day, almost entirely by sliding. Such a high sliding rate occurs because the glacier, by terminating in the ocean, must have a subglacial water pressure high enough to drive water out of the glacier against the pressure of the ocean water.

Glacier hydrology. A temperate glacier is essentially a reservoir that gains precipitation in both liquid and solid form, stores a large share of this precipitation, and then releases it with little loss at a later date. The hydrologic characteristics of this reservoir, however, are complex, because its physical attributes change during a year.

In late spring the glacier is covered by a thick snowpack at the melting temperature. Meltwater and liquid precipitation must travel through the snowpack by slow percolation until reaching well-defined meltwater channels in the solid ice below. In summer the snowpack becomes thinner, and drainage paths within the snow are more defined, so that meltwater and liquid precipitation are transmitted through the glacier rapidly. In winter, snow accumulates, and the surface layer freezes, stopping the movement of meltwater and precipitation at the surface. The rest of the ice reservoir may continue to drain, but in the process the conduits within and under the ice tend to close.

Runoff from glacier ice

The runoff from a typical Northern Hemisphere temperate glacier reaches a peak in late July or early August. Solar radiation, the chief source of heat to promote melt, reaches a peak in June. The delay in the peak melt rates is primarily because of the changing albedo (surface reflectivity) during the summer; initially the snow is very reflective and covers the whole glacier, but as the summer wears on the snow becomes wet (less reflective), and in addition more and more ice of much lower albedo is exposed. Thus, even though the incoming radiation decreases during midsummer, the proportion of it that is absorbed to cause melt is greatly increased. Other heat-exchange processes, such as turbulent transfer from warm air, also become more important during midsummer and late summer.

This albedo variation produces a runoff "buffering effect" against unusually wet or dry years. An unusually heavy winter snowpack causes high-albedo snow to persist longer over the glacier in summer; thus, less meltwater is produced. Conversely, an unusually light winter snowfall causes older firn and ice of lower albedo to be exposed earlier in the summer, producing increased melt and runoff. Thus, glaciers naturally regulate the runoff, seasonally and from year to year. When glacier runoff is combined with nonglacier runoff in roughly equal amounts, the result is very stable and even streamflow. This condition is part of the basis for the extensive hydrologic development that is found in regions such as the Alps, Norway, and western Washington.

Glacier streams are characterized by high sediment concentrations. The sediment ranges from boulders to a distinctive fine-grained material called rock flour, or glacier flour, which is colloidal in size (often less than one micrometre in diameter). The suspended sediment concentration decreases with distance from the glacier, but the rock-flour component may persist for great distances and remain suspended in lakes for many years; it is responsible for the green colour of Alpine lakes. Glacier streams vary in discharge with the time of day, and this variation causes a continual readjustment of the stream channel and the transportation of reworked debris, adding to the sediment load. Rates of glacier erosion (that is, sediment production) are typically on the order of one millimetre per year, averaged over the glacier area, but they are higher in particularly steep terrain or where the bedrock is especially soft.

The *jökulhlaup*

Glacier floods. Glacier outburst floods, or *jökulhlaups*, can be spectacular or even catastrophic. These happen when drainage within a glacier is blocked by internal plastic flow and water is stored in or behind the glacier. The water eventually finds a narrow path to trickle out. This movement will cause the path to be enlarged by melting, causing faster flow, more melting, a larger conduit, and so on until all the water is released quite suddenly. The word *jökulhlaup* is Icelandic in origin, and Iceland has experienced some of the world's most spectacular outburst floods. The 1922 Grimsvötn outburst released about 7.1 cubic kilometres (1.7 cubic miles) of water in a flood that was estimated to have reached almost 57,000 cubic metres (2,000,000 cubic feet) per second. Outburst floods occur in many glacier-covered mountain ranges; some break out regularly each year, some at intervals of two or more

years, and some are completely irregular and impossible to predict.

Glacier surges. Most glaciers follow a regular and nonspectacular pattern of advance and retreat in response to a varying climate. A very different behaviour pattern has been reported for glaciers in certain, but not all, areas. Such glaciers may, after a period of normal flow, or quiescence, lasting 10 to 100 or more years, suddenly begin to flow very rapidly, to up to five metres per hour. This rapid flow, lasting only a year or two, causes a sudden depletion of the upper part of the glacier, accompanied by a swelling and advance of the lower part, although these usually do not reach positions beyond the limits of previous surges. Advances of several kilometres in as many months have been recorded. Even more interesting is the fact that these glaciers periodically repeat cycles of quiescence and activity, irrespective of climate. These unusual glaciers are called surging glaciers.

Although surging glaciers are not rare in some areas (*e.g.*, Alaska Range and St. Elias Mountains), they are totally absent in other areas of similar topography, bedrock, climate, and so forth (*e.g.*, western Chugach Mountains and Coast Mountains). Furthermore, glaciers of all shapes and sizes, from tiny cirque glaciers to major portions of a large ice cap, have been known to surge. The flow instability that results in glacier surges is generally caused by an abrupt decoupling of the glacier from its bed. This decoupling is the result of a breakdown in the normal subglacier water flow system, but the exact mechanisms that cause some glaciers to surge are not fully understood.

Difficulty of explaining surging glaciers

Tidewater glaciers. Many glaciers terminate in the ocean with the calving of icebergs. Known as tidewater glaciers, these glaciers are the seaward extensions of ice streams originating in ice fields, ice caps, or ice sheets. Some tidewater glaciers are similar to surging glaciers in that they flow at high speeds—as much as 20 to 25 metres per day—but they do so continuously. Tidewater glaciers share another characteristic with surging glaciers in that they may advance and retreat periodically, independent of climatic variation.

The physical mechanisms that control the rate of iceberg calving are not yet well understood. Empirical studies of grounded (not floating) tidewater glaciers in Alaska, Svalbard, and elsewhere suggest that the speed of iceberg calving is roughly proportional to water depth at the terminus. This relation can produce an instability and periodic advance-retreat cycles. For example, a glacier terminating in shallow water at the head of a fjord will have a low calving speed that may be exceeded by the ice flow speed, causing advance of the terminus. At the same time, glacial erosion will cause the deposition of sediment as a moraine shoal at the terminus. With time, the glacier will advance, eroding the shoal on the upstream face and depositing sediment on the downstream face. The shoal, by reducing the depth of the water at the glacier's terminus and thereby inhibiting iceberg calving, will allow the glacier to advance into deep water farther down the fjord. This advance phase is slow—typically 10 to 40 metres per year—and in an Alaskan fjord it may take a period of 1,000 years or more to cover a typical fjord length of 30 to 130 kilometres.

Such a glacier, in an extended position and terminating in shallow water on a moraine shoal, is in an unstable situation. If, for some reason, the terminus retreats slightly, the deeper water upstream of the shoal will cause an increase in iceberg calving; this will result in further retreat into deeper water, which will further increase the calving until the calving speed becomes so high that the normal processes of glacier flow cannot compensate. A rapid, irreversible retreat will result until the glacier reaches shallow water back at the head of the fjord. In contrast to the slow advance phase, the retreat phase may take only a few decades. The fastest glacier retreats observed during historical time (for instance, the opening of Glacier Bay, Alaska), as well as those inferred during the demise of the great Quaternary ice sheets, were caused by this mechanism. Information on the advance and retreat of tidewater glaciers should not be used to infer climatic change, however. (M.F.M.)

Icebergs and pack ice

Ice in the waters of the Earth's polar regions occurs in two forms—namely, pack ice and icebergs. Pack ice forms from seawater and is generally only one to two years old, whereas icebergs are fragments of ice sheets and glaciers that formed on land areas during intervals of thousands of years. Pack ice expands during winter to cover large areas of the oceans in both hemispheres. Melting occurs in spring and summer, and the margins of the pack ice retreat. This warmer weather aids calving (separation) of icebergs at the boundaries of ice sheets and glaciers, however, and many icebergs start their transit toward the Equator at this time.

Importance of ice at sea

This ice at sea is of substantial importance to humans. Major shipping routes around the world could be shortened by as much as 30 percent if regular navigation through the Arctic Ocean became feasible, for example. The economics of freshwater extraction from icebergs also is instructive. An iceberg 16 kilometres (10 miles) long could be hauled into the Peru (Humboldt) Current, which flows from the south along South America, and ultimately to Los Angeles where it could supply fresh water to that city for three weeks. Factors of wind effect, breakup, grounding, accessible supply, and controlled melting on delivery argue against this venture, however. In the Northern Hemisphere about 10,000 icebergs are produced each year from the West Greenland glaciers and an average of 375 flow south of Newfoundland into the North Atlantic shipping lanes.

Because ice reflects four to five times more sunlight than does the ocean, the change in areal coverage of sea ice between winter and summer is climatically important. There is a change in areal coverage of 20 percent from winter to summer in Arctic pack ice extent, and 80 percent in Antarctic pack ice maximum to minimum coverage. This winter to summer change in area of pack ice coverage is equivalent to twice the area of the United States.

ICEBERGS

Formation and distribution. *Calving of glaciers and ice sheets.* Probably the first mention of icebergs was that of St. Brendan, an Irish monk whose partly fictional writings suggest that he encountered a "floating crystal castle" on the high seas. After calving (breaking off) from the Greenland and Antarctic ice sheets and smaller outlying glaciers, icebergs can move thousands of miles in a few years; those in the Arctic can slip down along the eastern North American coast to be caught up in the Gulf Stream and, while melting, be carried in a few weeks to within several hundred kilometres of England and Ireland or by a combination of wind and current to Bermuda, as happened in 1907 and 1926. These and other rare sightings are shown Figure 13.

Gulf Stream transport

As previously noted, when snow and freezing rain continue to precipitate over a continent in excess of evaporation, glaciers will form and icebergs will break off their ends and appear in the surrounding ocean. At the point where glaciers or large, extensive ice shelves meet the sea, water pressure beneath the ice shelf or glacier tongue interacts with the outward creeping glacier. The tides, which have ranges up to 6 metres (20 feet) in the Arctic, along with small sea-level changes associated with wind and swells, result in an intermittent increase and decrease in force on the protruding end of the glacier or ice shelf resulting in the birth of a large monolith of drifting ice. There are other ways in which an iceberg is formed. One, which is characteristic of southern Greenland glaciers, consists of a melting or evaporation of the surface portions of the glacier near its terminus at a greater rate than the water erosion on its underside. This results in an underwater shelf, and eventually, through the erosion of water and periodic tidal and other hydraulic forces, this is broken off and an iceberg floats to the surface. Icebergs of varying shapes are produced in this way. A third mechanism by which icebergs are formed is through gradual break-off from a hanging glacier or ice shelf. The type of iceberg mechanism is related to the surrounding topography, the climate, and the rate of flow of associated glaciers.



Figure 13: Limits of sea ice and icebergs in the Northern Hemisphere.

Movement of glaciers producing icebergs

The speed of the ice sheet flow, or creep, over Greenland and the Antarctic varies from zero near their centres to as much as 10 kilometres per year in the ice streams that make up glaciers. For the same inclination relative to gravity, ice moves 1/10,000 as fast as water. The average movement in the Antarctic is 360 metres per year, with most of the measurements between extremes of 110 and 1,100 metres per year. These measurements are made near the coast and are much higher than the average flow rate over the entire Greenland or Antarctic ice sheets.

Arctic icebergs. The western shore of Greenland has the fastest-flowing glaciers on the Earth. The one known as the Quarayaq Glacier flows at a velocity between 20 and 24 metres per day, greater than the velocity of most Alpine glaciers. Jakobshavn Glacier at latitude 70° N, approximately, produces 10 percent of all the Greenland icebergs (approximately 1,350 annually) and flows at about 20 metres per day. The icebergs accumulate in a fjord and periodically spill from this fjord in groups accompanied by noise that can be heard for several kilometres. This greatest iceberg-producing glacier measures only 7 kilometres along its front and is 90 metres above sea level. In contrast to the rapidly moving glaciers, the very wide glaciers that move slowly produce only very small icebergs. Some glaciers, such as the Frederikshåb, Greenland, with a 33-kilometre front, have rates of flow equal to the rate of melting, and thus produce no icebergs. The annual yield of icebergs in the Arctic is, at the most, 15,000, with only 5,000 or so of sufficient size to reach the open ocean intact.

The largest glacier in the Northern Hemisphere is the Petterman Glacier, at 81° N 62° W. Although it is only a few metres above seawater, its ice foot extends as far as 40 kilometres out to sea, and it pushes a path through old piled-up sea ice. These long fingers of glacier ice, under severe climatic conditions, break off once every 10 to 20 years. Another glacier of importance is the Jungersen Glacier in northern Greenland. This glacier and the Petterman Glacier produce very large tabular icebergs, known in the Arctic as "ice islands." They are similar in shape and mode of formation to the large tabular icebergs of the Antarctic but are much smaller.

East Greenland icebergs tend to move northward; they are small and few in number. The icebergs that do leave the fjord and the coast area enter the East Greenland Current, and some join the West Greenland icebergs. The icebergs that reach the North Atlantic Ocean from northern Greenland, Siberia, and the Northwest Territories constitute less than 10 percent of the total icebergs produced in the Northern Hemisphere. They do not affect the sea routes and, other than their formation of ice islands, have no importance to humans. Most icebergs originating from western Greenland are of great importance; of an annual production of about 7,500 bergs (Arctic total is 10,000–15,000), the Labrador Current carries 800 to 1,000 into the open ocean.

Sightings in the North Atlantic

Of the 10,000 to 15,000 icebergs calved from glaciers annually in the Arctic, only 375 on the average pass Newfoundland, or latitude 48° N, into the North Atlantic Ocean. In some years more than 1,000 are seen, in others fewer than 30. The yearly average diminished over 20 years until 1972, when 1,400 icebergs were sighted south of 48° N in the North Atlantic. The distribution of these icebergs in April, May, June, and July is shown in Figure 13.

There are few icebergs in the Arctic Basin proper. The major source area for icebergs in the Barents Sea is Franz Josef Land. Icebergs are not found in the North Pacific except in sounds along the Alaskan-Canadian coast between latitudes 55° and 60° N.

Antarctic icebergs. In the Southern Hemisphere the maximum limit of iceberg drift is 1,600 kilometres farther north than the northern extent of sea ice in the Antarctic (Figure 14). Most icebergs are concentrated south of the Antarctic current convergence at about 60° S. Some icebergs from the ice shelves of Antarctica drift north of latitude 42° S in the Atlantic Ocean but only to latitude 56° S in the South Pacific Ocean. One Antarctic iceberg was sighted only 50 kilometres south of the Cape of Good Hope (Africa) in 1850. The northern limit of ice sighted in the Southern Hemisphere was 26°30′ S.

Size of icebergs. *Arctic.* Arctic icebergs vary in size from the size of a large piano, called growlers, to the dimensions of a 10-story building. Icebergs about the size of a small house are called bergy bits. Many icebergs in the Arctic are about 45 metres tall and 180 metres long. Icebergs of the Antarctic not only are far more abundant but are of enormous dimensions compared with those in the Arctic. Ninety-three percent of the world's mass of icebergs is found surrounding the Antarctic.

Other than Arctic Basin ice islands, the largest iceberg in the Northern Hemisphere was 11 kilometres long and 5.9 kilometres wide; it was sighted near Baffin Land in 1882. The largest Arctic iceberg sighted south of Newfoundland was encountered by a convoy of ships during World War II at 43°10′ N and 49°33′ W. There were multiple collisions and much confusion before all ships safely limped into port. This ice island was 1,370 metres long, 1,100 metres wide, and 18 metres high. These icebergs are similar in size to Antarctic icebergs, but by and large only a few large tabular icebergs are seen in the Arctic as compared with the predominance of the tabular icebergs noted in the Antarctic. The tallest icebergs known were measured at 134 and 158 metres high; the latter was measured and photographed from a helicopter in the late 1950s.

Antarctic. Antarctic icebergs are characterized by their tremendous size and tabular shape. Lengths up to seven kilometres are not unusual, with ice 45 metres above water. The discovery of the origin of these immense tabular bergs was made in 1841 when the Ross Sea was penetrated and the Ross Ice Shelf was discovered to be afloat. Most Antarctic icebergs are formed from the Antarctic continental ice sheet as it thins toward the coast and exudes into the ocean as a great ice shelf with fronts hundreds of kilometres long. The four major ice shelves are the Ronne-Filchner Shelf in the Weddell Sea, the Ross Ice Shelf in the Ross Sea, the Shackleton Ice Shelf in the Indian Ocean sector, and the Larsen Ice Shelf on the Antarctic Peninsula (also known as Palmer Land and Graham Land). Antarctic icebergs, if they remain locked in the pack ice, will last for many years. One of the largest icebergs sighted was over 140 kilometres in length. This tabular iceberg was first sighted in 1927, and presumably the same iceberg was later seen in 1931, at which time it was 100 kilometres long. The largest known Antarctic iceberg was measured by the icebreaker USS *Glacier* in 1956; it had a length of 333 kilometres and a width of 100 kilometres.

Age and melting. Greenland icebergs 300 to 450 metres thick represent several centuries of precipitation. This figure is based on the comparison of the thickness of the ice sheet to the known or average annual precipitation of 20 to 60 centimetres (8 to 24 inches) per year in the source area for icebergs. Age of ice in central Greenland, close to bedrock, is estimated at 30,000 to 150,000 years old. The longest core retrieved from the Greenland ice sheet was 1,370 metres long with a bottom date of 100,000 years. It is probable that the oldest ice melts before reaching the outlet glaciers. From the known amount of precipitation over Greenland, and the assumption that the ice volume and precipitation rate now are approximately the same as they were many centuries ago, it can be estimated that the mean age of icebergs is 5,000 years. Measurements by carbon-14 dating of entrapped air show that icebergs are hundreds to thousands of years old—the oldest actual measurement being 3,000 years. Arctic ice islands and giant Antarctic bergs last as long as 10 years at high latitude. Most western Greenland icebergs melt within two years.

Deterioration of icebergs at sea

Once an Arctic iceberg has been calved and moves out to the open sea, it sojourns in Baffin Bay for three months to two years, during which time it undergoes some disintegration through melting and calving of small chunks of ice from its perimeter. This results in a decrease in mass of about 90 percent by the time it reaches the coast of Newfoundland and the Grand Banks in the North Atlantic. When the iceberg enters the region of the Grand Banks, where the warm waters of the Gulf Stream meet the colder waters of the Labrador Current, it has only a few days of life remaining. A large iceberg 120 metres long melted within 36 hours in 27° C (80° F) water. The estimated rate of iceberg melting is based on the observations of a number of individuals from the International Ice Patrol. For mild sea conditions an iceberg deteriorates at a rate of height decrease of two metres per day in 0° to 4° C (32° to 39° F) water, and three metres per day in 4° to 10° C (39° to 50° F) water. Destruction of icebergs in warm water is increased during stormy weather, when mechanical erosion of icebergs is added to the thermal effects of air and water. During the erosion process icebergs usually take on the form of a saddle, because erosion at one pole of the major axis of the iceberg results in that point rising, while the other end of the major axis is being eroded. Subsequently the latter end, owing to loss in weight, arises, and this rocking back and forth continues while constant erosion is occurring along the minor axis leading, usually, to a bipeaked or saddle-shaped structure. Table 2 is an estimate of the time necessary for deterioration and is based on unpublished quantitative measurements in 1960 and general observations by ice scientists.

Table 2: Representative Deterioration or Melting Times for Icebergs
(at 45° N; estimated)

seawater temperature		melting time (in days)	
°F	°C	*	†
32	0	40	80
40	4	10	20
70	21	4	8

*For an iceberg 80 feet high, 300 feet long.
†For an iceberg greater than 150 feet high, 500 feet long.

Wind and current effects. *Drift of icebergs.* Iceberg movement is influenced by direct wind push on its exposed area to an extent far greater than commonly assumed. Although the bulk of the iceberg is below water, in many situations wind has a dominant influence on the movement. The wind intensity and direction over Baffin Bay in the spring of one year influences the number of icebergs that slip into the North Atlantic that year and the following year. This can be understood by noting that icebergs pouring out of the Arctic north of Newfoundland in spring will run aground or be trapped in the embayments of western Baffin Bay unless wind and current deflect them to the southeast. Of more importance is the effect of wind over western Greenland, however, where intense offshore early summer winds over the ice fjords drive sea ice-entrapped bergs into the West Greenland Current, thus increasing the number of icebergs that, having made the usual counterclockwise circuit, will arrive off Newfoundland the following spring. There is a reasonable correlation between the atmospheric pressure distribution one year and the number of icebergs drifting south of Newfoundland in the following year.

Table 3: Exposed-to-Submerged Proportions, and Wind Factors, for Icebergs

iceberg description	proportions (exposed : underwater)	wind factor
Flat-topped or blocky	1 : 6	.004
Rounded or domed	1 : 4	.005
Usual Greenland iceberg	1 : 3	.01
Pinnacled or drydock	1 : 2	.03
Winged	1 : 1	.04

Effect of size and shape

The day-to-day movement of an iceberg is controlled by the size and shape of the iceberg, previous and present wind, surface wind current, and general ocean current. The most important factor in assessing wind drift of icebergs is size and shape. Although most icebergs have a specific gravity of 0.9, and thus $^9/_{10}$ of the mass is below the sea surface, it is not true that this means that a 30-metre-high iceberg is 180 metres deep in all cases. This is true only for the rectangular, blocky, or flat-topped icebergs common to the Antarctic. Table 3 gives the relation between exposed and underwater areas for various shapes of icebergs.

Winged icebergs, those with saillike pinnacles around the central mass, are very much influenced by the winds and move at speeds of 1 knot, or 24 nautical miles per day, under the influence of steady winds of 30 knots. The wind force on an iceberg does not result in movement directly downwind, but, because of the rotation of the Earth (Coriolis effect), windage on an iceberg is 30° to 50° to the right in the Northern Hemisphere and to the left in the Southern Hemisphere.

Windage and drift

The momentum of icebergs is so great that once in motion they continue for hours after the wind has abated. It is possible for an iceberg to be driven before a 30-knot wind at 1 knot in a direction 30° to the right of the wind, and after the wind stops the iceberg will slow and circle in what is known as an inertial circle, associated with the rotation of the Earth. An iceberg near the Grand Banks will be moving in a direction opposite to that toward which the wind was blowing about eight hours after the wind stops; however, the speed would be low and such anomalous movement would not be observed unless the initial speed and momentum were great. A careful accounting of berg underbody dimensions and the relationship of wind-current forces and iceberg mass will lead to explanations of anomalous iceberg movements. To emphasize the importance of wind effect and its prediction, one potentially disastrous case will be cited. In 1960 a 60-metre-high, 300-metre-long iceberg was driven off the tail of the Grand Banks into the shipping lanes by northwesterly winds averaging 80 kilometres per hour. This iceberg moved 140 kilometres at as much as 3 knots across the Labrador Current and resulted in an emergency move of the North Atlantic shipping lanes to the south.

The drift of icebergs and pack ice is the result of the ocean current and the wind. When the winds are variable or less than 32 kilometres per hour and the current greater than 0.5 knot, the current predominates, but, when steady 30-knot winds blow for more than 12 hours, the wind effect becomes important even in areas where the ocean current is 1 to 2 knots. In addition to windage on the iceberg and the ocean gradient current, the wind-induced surface current has the effect of increasing drift speed by about 10 percent for small icebergs and increasing the angle of drift direction.

Drift of sea ice. Sea ice drift is better understood, but in some respects it is more complicated than iceberg drift. In addition to size, inertia, and exposed-to-underwater dimensions, important additional information on surface roughness, water drag, and internal resistance due to ice-field concentration is needed to describe the motion. The observations of American and Russian scientists drifting on ice islands in the Arctic Basin, Baffin Bay, and Gulf of St. Lawrence, along with Japanese and Russian long-term ice-field observations, are summarized in Table 4. As noted from the table, ice fields consisting of 10 percent ice and 90 percent open water will move at 1 to 8 percent of the surface wind velocity. The angle of drift is from 20° to 40° to the right of the wind in the Northern Hemisphere. The speed will be reduced by a factor of four as the ice becomes packed to 90 percent coverage of the ocean. Smooth ice drifts with less speed than rough ice because the wind has greater effect on a rough surface. Rough or hummocked ice usually is thicker and thus has greater inertia; ice of great inertia takes longer to reach the wind factor speed, but, after the wind stops, the ice continues to move longer than light ice. This phenomenon results in strings of ice floes aligned perpendicular to the wind with small floes packed to windward against larger floes. This wind sorting of ice is a phenomenon of great importance in navigating pack ice, and the strips of open water aligned in a direction approximately perpendicular to the direction of wind explain the successful navigation in ice performed by sailing vessels in the past.

Sediment transport. Both icebergs and pack ice transport sediment in the form of pebbles, cobbles, boulders, and finer material, and even plant and animal life, thousands of miles from their source area. The distribution of icebergs 10,000 years ago, for example, can be inferred from sediments that are widely disseminated on the ocean floor of the North Pacific Ocean as far south as latitude 48° N and in the South Atlantic Ocean as far north as the latitude of the Cape of Good Hope in the Southern Hemisphere. Ice rafting competes with kelp rafting as an explanation for the occurrence of pebbles and boulders on the seafloor as far south as Baja California and other areas of the world ocean.

Bottom freezing of ice shelves and reduction of the ice surface results in migration of sediments and organisms upward. Layering of sediment can be seen in sea-ice floes and icebergs from ice shelves. Fossil penguin bones have been found as far north as 30° S, and massive banks of boulders have been noted on the seafloor off South Africa. Thicknesses of as much as 130 centimetres of glacial till have been noted on the ocean floor from discharge by icebergs, apparently during the past one million years.

Icebergs are coloured brown, black, and green by a combination of sediment, plankton deposits under the source area ice shelf, and glacial blue ice.

Iceberg detection and destruction. In the open ocean most ice is seen by radar at ranges depending on fragment size, but smaller icebergs or growlers can be detected only when the sea surface is calm, and then only at ranges of about two kilometres. During slight wind conditions, in particular in heavy seas, the echoes from the waves, known as sea return, may completely mask the echoes of large, potentially very hazardous chunks of ice. In addition, radar return from rain and snow obscures the return from ice in either rain, snow, or fog conditions. Neither radar nor sonar can be relied upon for detection of icebergs and pack ice in choppy seas. The reflectivity of ice and snow to light is great, but reflectivity to radar or short radiowaves is very poor. An iceberg seven metres high cannot be detected with modern equipment if the waves are over one metre high. Various innovations since the inception of radar during World War II have not improved significantly the capabilities of radar to detect icebergs; thus ships are still bound by international agreement to proceed at slow speed in fog.

Reliability of radar detection

Sonar is effective in detecting icebergs; however, the range of detection is frequently limited by the water conditions and speed of commercial vessels. The likelihood of insufficient warning for high-speed passenger and cargo ships leaves this mode of detection inadequate. Other sug-

Table 4: Wind Factor for Speed of Sea Ice Drift

sea covered by ice (percent)	wind factor	
	rough hummocky surface	smooth floes
10	.08	.01
50	.05	.005
90	.02	.003

gestions of dropping radio transducers or metal reflectors onto icebergs have the common fault that they are subject to icebergs rolling over and calving, which are frequent occurrences as a dying iceberg melts its way deep into the warmer waters of the shipping lanes.

The problem of iceberg protection, therefore, is one of tracking icebergs as they come down the Labrador Current and reporting the whereabouts of these floating menaces to all North Atlantic shipping as often as twice daily. This sometimes involves 300 icebergs and requires a team of iceberg experts, oceanographers, aviators, and seamen. During heavy ice conditions two U.S. Coast Guard planes fly six- to eight-hour reconnaissance missions from Argentia, Nfd., Can., over the Grand Banks off Newfoundland and contiguous areas. Positions of icebergs are plotted and are correlated with previous positions of the same icebergs or groups of icebergs. When dangerous icebergs approach the shipping lanes, a ship departs for the scene to stand by near the iceberg and warn approaching ships.

Once an iceberg is spotted in a position of threat to ships, it should be of no harm if destroyed. Destroying a 200,000-ton block of ice is, however, a task whose difficulties leave it impractical in most situations. The first successful results on breaking up icebergs by explosives were reported in 1929, when a few hundred kilograms of thermite cracked an iceberg into smaller pieces through thermal stress; these then melted more rapidly owing to the greater surface area exposed. Similar thermite experiments on two icebergs in 1959 did not corroborate the original findings, however. Attempts at bombing, torpedoing, shelling, and even ramming have been unsuccessful. With twenty 450-kilogram (1,000-pound) bombs acting as direct hits, or as depth charges, it is possible to chip away about 20 percent of a 250,000-ton iceberg. The International Ice Patrol has even tried painting half of an iceberg with lampblack or charcoal to induce thermal stress by heat absorption, but, like other experiments, these attempts were inconclusive.

PACK ICE

Formation and characteristics. On superficial examination, frozen ocean or salt water appears similar to freshwater ice; there are two principal differences, however. First, because the maximum density of seawater occurs below the freezing point, even after freezing the water below the ice will continue to turn over or circulate. As the surface water near the ice becomes colder, it becomes heavier and sinks, resulting in a continuous turnover or vertical circulation of the water beneath the ice. This is a different situation from that which occurs in lakes where, because the maximum density of fresh water is above the freezing point, once ice forms the colder water is lighter than the deeper, somewhat warmer water, and mixing does not occur (see above *Ice in lakes and rivers: Ice in lakes*). A second characteristic is the fact that, as seawater freezes, minute pools of salty water called brine pockets are entrapped. The final form and the macroscopic physical properties of the ice are very much dependent upon the concentration of the brine pockets within the ice block. Once an ice field has formed, the physical and chemical properties are not locked in a frozen coffin but vary as brine pockets migrate through the ice block in response to gravity and thermal gradients.

Crystallization. In the Northern Hemisphere during September and October, the air temperature lowers sufficiently to form a thin sheet of ice. Freezing temperature for average northern ocean salt water of about 3.5 percent salt composition by weight (usually designated 35 parts per thousand) is −1.8° C (28.8° F). The first signs of freezing are changes in the colour and texture of the sea surface as thin, gray-coloured needles and crystal plates form a surface-thin sludge. If quiet sea conditions prevail, sheets of crystalline aggregates plate the ocean surface. Initially the ice film is entirely fresh, but, as more ice crystals form, pockets of salt water (brine pockets) become entrapped between lamellae (very fine layers) of tiny ice plates. The amount of brine entrapped depends on the temperature of formation and the age of the ice. The shape of the initial ice crystals varies from square discoids to hexagonal dendritic forms. The average width is 2.54 centimetres (1 inch) and the thickness up to about 0.15 centimetre (0.06 inch). During the surface veneer formation stage each grain is free to grow both laterally and vertically until the sheet consolidates. As the ice sheet thickens, the orientation of these grains will change. Owing to slight breezes and water motion, the thin sheets of ice jostle about and, after but a few hours, form a field of ice paddies.

The freezing process

The appearance is very much similar to a lily pond completely covered with large gray lily leaves with slightly raised white fringes around their periphery. These disks of ice are known as pancake ice. If the temperature remains below freezing, the pancake ice coalesces as more ice forms, and within a few days the ice cover can be about 8 to 10 centimetres thick with a slightly corrugated surface, unless snow prevails, in which case the entire sea area appears as a smooth white plain. As seawater continues to freeze at the bottom edge and sides of ice floes and fields, snow cover increases and the pressures associated with the stresses and strains caused by water and wind movement result in a hummocking and ridge development in some places and open water in other places.

Growth of the ice sheet. The rate at which the ice forms and thickens depends on the air temperature, ocean turbulent heat flux (mixing conditions), and amount of snow acting as a heat or cold insulator. An empirical formula has been developed and used extensively by Russia and the United States to predict ice appearance and growth rate. The equations are based on the simple concept that ice growth is directly related to the length of time during which the air temperature is below the freezing point for seawater. By adding the number of degrees below the freezing point for each day, a measure of the severity of cold and time of exposure is obtained. This measure is known as the freezing degree days. In north polar regions there are about 8,000 freezing degree days, which is equivalent to four months of −18° C (0° F) air temperature or a mean annual Arctic air temperature of −12° C (10° F; 22° F below freezing).

During ice growth there is surface evaporation and sublimation and bottom ablation when upward heat conduction in the ice is less than ocean upward heat conduction. The balance between ablation and freezing or accumulation results in an equilibrium thickness of about 3.5 metres (11.6 feet) of ice in the Arctic and probably about the same in the Antarctic. The lifetime of this North Polar sea ice is five to eight years, and the lifetime of Antarctic polar class ice found only in the Bellingshausen and Weddell seas is about three years. These lifetime values are related to the rate with which the whole Arctic or Antarctic pack in certain areas moves toward the Equator.

Sometimes early in the season there is sufficient warming and wind-induced surface motion to completely disintegrate the ice field. Oftentimes after about 10 to 13 centimetres of ice and snow have been formed in a more-or-less uniform manner, a large crack develops, which, through wind and stress, opens into a wide canal commonly known as a lead. This phenomenon is seen frequently in older ice and during the spring breakup. Leads are frequently followed by ships navigating in ice fields, but unfortunately the leads sometimes have a dead end with a very large iceberg blocking the way. With alternating freezing, partial melting, snow, and wind and swell, the ice field develops over a matter of a few weeks to a month into a 15- to 61-centimetre-deep ocean cover. At this point the ice field is still navigable by most large vessels; however, if a vessel finds itself in the far north two weeks after the commencement of active surface freezing (*e.g.*, in late October), it is in peril of being locked in for the remainder of the winter.

Saltiness of sea ice

Salt content. The salt content of seawater as it freezes is always less than that of normal seawater. The amount of salt in the seawater during its first moment in the solid state is dependent upon the rapidity with which the seawater freezes, but in general the salt content is about one-tenth that of seawater. The slower the freezing process, the less the salt content. The most rapid freezing is that which occurs during the first day, and this ice is saltier than underlying ice which forms at the ice-water interface. The salt that remains in the ice is located in tiny pockets of fluid surrounded by normal crystals. These pockets of

fluid migrate, mainly by gravity, through the matrix of ice crystals with the result that, after a few weeks to a few months, the surface of the ice becomes lower in salt content than the deeper layers. It had once been thought that the difference in temperature, or thermal gradient, was the principal driving force for brine pocket migration; careful experiments indicate, however, that although the thermal gradients are a factor, it is principally gravity that accounts for the movement of these brine pockets. This migration continues throughout the winter.

In the summer, when the ice temperature rises, there is a rapid increase in the migration of salt out of the ice. The sea ice at the surface loses so much salt that it becomes potable and, in fact, is used by Eskimos as a source of fresh water. Salinity reaches a value of less than 0.01 percent. In summary, when first formed, the surface layer may have a salinity of 2 to 4 percent, but by April the salinity has dropped to between 0.4 and 0.7 percent, while sea ice that has been through at least one melt season has a salt content below 0.1 percent.

Arctic pack ice. *North Atlantic.* The pack ice of the Northern Hemisphere covers an average area of 10,620,000 square kilometres (see Figure 13), filling the Arctic Ocean basin and adjacent North Atlantic Ocean. The polar ice field consists of 4,700,000 square kilometres of three- to six-metre-thick polar ice that never melts. Infrared imagery from aircraft, however, shows that 10 percent of the polar pack is open water even during winter.

Ice streams into the North Atlantic

Along with Arctic Basin seasonal sea ice, this Arctic pack exudes into the northern Atlantic through two ice streams. The major exit of drifting pack ice from the Arctic Basin is along the eastern side of Greenland, mostly west of Spitsbergen. This ice tongue stretches 2,400 kilometres out of the Arctic Ocean and empties a stream of sea ice at a drift rate of almost 13 kilometres per day. The second icy arm of the north consists of a discharge through the Arctic-Canadian Archipelago and along the eastern American shore. This outpouring of ice is the principal deterrent to easy Northwest Passage ship transit and northern American migration and exploration. During winter, fast ice and local sea ice form along the Siberian coast, Barents and Kara seas, East Greenland, and Labrador coasts down to Newfoundland. The maximum extent of drifting sea ice is approximately latitude 42° N (about the same latitude as that of Boston); however, this represents the limit of floating ice pieces and not the hazardous ice-pack edge, which seldom reaches south of Newfoundland. During the summer the 240-kilometre belt of ice lying along the Labrador coast from Newfoundland northward melts to leave the approaches into Hudson Bay and the Canadian Northwest Territories clear. The motion of the polar ice follows an enormous clockwise eddy with a centre 85° N 170° W.

North Pacific. In the North Pacific Ocean comparatively little pack ice and icebergs are encountered. The Bering Sea is clear of pack ice during the northern summer, but commencing in September pack ice forms in bays and is carried through the Bering Strait. In winter and spring pack ice is found as far south as 40° N. This is drifting ice from northern latitudes and the Sea of Okhotsk. During winter, pack ice forms in the northern part of the Sea of Japan.

Antarctic pack ice. Approximately twice as much pack ice forms in the oceans surrounding Antarctica as is found in the Arctic. There are, however, only limited regions in the Bellingshausen and Weddell seas where true polar perennial ice similar to the polar ice cap of the Arctic occurs. The maximum area of Antarctic pack ice is 20 million square kilometres, or about 8 percent of the Southern Hemisphere.

Figure 14 indicates the extent of Antarctic pack ice. It forms a fairly constant band of drifting sea ice around the continent, with the farthest northern extent occurring at the end of the austral (southern) winter in October. The greatest extension of pack ice in the South Pacific sector is found in about latitude 62° S, and in the South Atlantic pack ice extends to roughly 52° S. The average northern boundary for icebergs is 56° S in the Pacific sector and 42° S in the South Atlantic. The minimum ice coverage occurs in March, when most of the Antarctic coast is free of ice, with the exception of the Weddell and Bellingshausen seas. The eastern and western coasts of the Weddell Sea are ice-free, but the Weddell itself is covered by a slowly (clockwise) revolving ice pack that seems to have a two-year cycle. The west coast of the Ross Sea is the most predictably open area during the Antarctic summer, and it is here at the approaches to McMurdo Station that most of the Antarctic expeditions have worked their way to the continent.

A substantial portion of the sea ice encountered in November and December on approaching the continent hundreds of kilometres from landfall represents ice that formed near the coast from the previous austral winter. In January and February there usually is clear water adjacent to the coast in all sectors around the continent, but this navigable water might be blocked by hundreds of kilometres of pack ice barrier farther out to sea. The Antarctic freeze-up commences with pack ice formation in the southerly parts of the Weddell Sea followed by pack ice appearance in the Bellingshausen and Ross seas. Beginning in March, ice is formed in sheltered bays, and it extends northward as the sea surface temperature drops. From late February to August, snowfall adds more to the ice thickness than is the case in the Arctic. This gives rise to an important difference between Arctic and Antarctic navigation, in that the presence of snow has a cushioning effect and ice breaking is more difficult. By the October maximum, the action of wind, sea, and some melting results in extremely active ice movement. It was in October that Sir Ernest Shackleton's ship *Endurance* was crushed and sank in the Weddell Sea in 1915.

The movement of the ice sheet around the Antarctic continent is from east to west except in the most northern part of the Weddell Sea, where there is a west to east movement making up the northern arm of the Weddell Sea Gyral. This clockwise Weddell eddy has been well documented by the drifts of entrapped ships. The ice drift in the Bellingshausen Sea is less definite. The ship *Antarctic* followed a meandering, aimless course when locked in the ice throughout the 1898–99 winter. From the Ross Sea, ice definitely drifts toward the Weddell Sea under the influence of the prevailing easterly winds near the Antarctic coast.

Sea ice forecasting and reconnaissance. Sea ice reconnaissance and forecasts in the Arctic and Antarctic are conducted by a number of nations, but the principal work is done by the U.S. Navy, either through the Fleet Weather Office or the Ice Forecasting Central of the U.S. Naval Oceanographic Office. The Ice Forecasting Central has for many years supported Arctic and Antarctic research and supply missions through ice reconnaissance and careful short- and long-range forecasting. Additional supportive reconnaissance and regional forecasts are provided by the Canadian government and U.S. Coast Guard ships. Support is even received from commercial airplanes, which frequently spot lonely icebergs far from the area of usual occurrence.

Satellite photographs lack the resolution for day-to-day iceberg reconnaissance; these photographs do show the edge of the sea ice in both polar regions, however. Ice navigation in the Antarctic is similar to navigation for logistic support of military and scientific expeditions in the Arctic. The emphasis is on pack ice distribution and concentration rather than icebergs, which offer little problem to ships once in the ice pack. (T.F.B./Ed.)

Permafrost

Permafrost, or perennially frozen ground, is naturally occurring material that has a temperature colder than 0° C (32° F) continuously for two or more years. This layer of frozen ground is designated exclusively on the basis of temperature. Part or all of its moisture may be unfrozen, depending on the chemical composition of the water or depression of the freezing point by capillary forces. Permafrost with saline soil moisture, for example, may be colder than 0° C for several years but would contain no ice and would not be firmly cemented. Most permafrost, however, is consolidated by ice; permafrost with no water, and

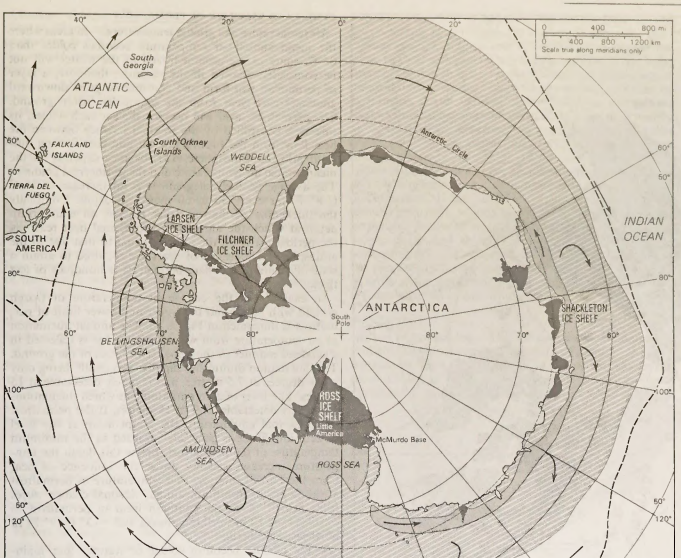


Figure 14: Limits of sea ice and icebergs in the Southern Hemisphere.

The active layer

thus no ice, is termed dry permafrost. The upper surface of permafrost is called the permafrost table. In permafrost areas the surface layer of ground that freezes in the winter (seasonally frozen ground) and thaws in summer is called the active layer. The thickness of the active layer depends mainly on the moisture content, varying from less than a foot in thickness in wet, organic sediments to several feet in well-drained gravels.

Permafrost forms and exists in a climate where the mean annual air temperature is 0° C (32° F) or colder. Such a climate is generally characterized by long, cold winters with little snow and short, relatively dry, cool summers. Permafrost, therefore, is widespread in the Arctic, sub-Arctic, and Antarctica. It is estimated to underlie 20 percent of the world's land surface.

DISTRIBUTION IN THE NORTHERN HEMISPHERE

Permafrost zones. Permafrost is widespread in the northern part of the Northern Hemisphere, where it occurs in 85 percent of Alaska, 55 percent of Russia and Canada, and probably all of Antarctica. Permafrost is more widespread and extends to greater depths in the north than in the south. It is 1,500 metres (5,000 feet) thick in northern Siberia, 740 metres thick in northern Alaska, and thins progressively toward the south.

Most permafrost can be differentiated into two broad zones; the continuous and the discontinuous, referring to the lateral continuity of permafrost (see Figure 15). In the continuous zone of the far north, permafrost is nearly everywhere present except under the lakes and rivers that do not freeze to the bottom. The discontinuous zone includes numerous permafrost-free areas that increase progressively in size and number from north to south. Near the southern boundary, only rare patches of permafrost have been found to exist.

Alpine permafrost

In addition to its widespread occurrence in the Arctic and subarctic areas of the Earth, permafrost also exists at lower latitudes in areas of high elevation. This type of perennially frozen ground is called Alpine permafrost. Although data from high plateaus and mountains are scarce, measurements taken below the active surface layer indicate zones where temperatures of 0° C or colder persist for two or more years. The largest area of Alpine permafrost is in western China, where 1,500,000 square kilometres (580,000 square miles) of permafrost are known to exist. In the contiguous United States, Alpine permafrost is limited to about 100,000 square kilometres in the high mountains of the west. Permafrost occurs at elevations as low as 2,500 metres in the northern states and at about 3,500 metres in Arizona.

A unique occurrence of permafrost—one that has no analogue on land—lies under the Arctic Ocean, on the northern continental shelves of North America and Eurasia. This is known as subsea or offshore permafrost.

Study of permafrost. Although the existence of permafrost had been known to the inhabitants of Siberia for centuries, scientists of the Western world did not take seriously the isolated reports of a great thickness of frozen ground existing under northern forest and grasslands until 1836. Then, Alexander Theodor von Middendorff mea-

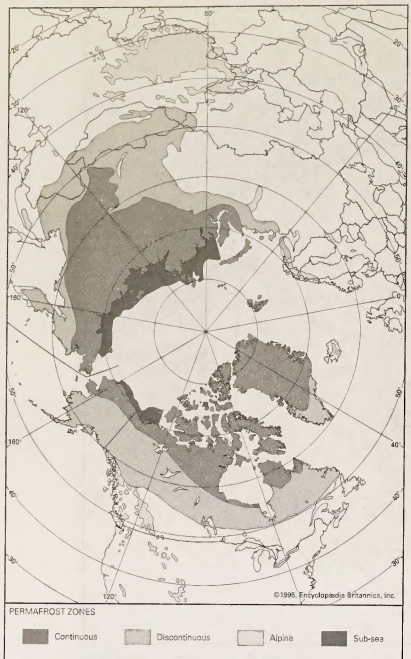


Figure 15: Distribution of permafrost in the Northern Hemisphere.
Adapted from T.L. Pewe, *Arctic and Alpine Research*, vol. 15 (1983), no. 2, p. 148

sured temperatures to depths of approximately 100 metres of permafrost in the Shargin shaft, an unsuccessful well dug for the governor of the Russian-Alaskan Trading Company, at Yakutsk, and estimated that the permafrost was 215 metres thick. Since the late 19th century, Russian scientists and engineers have actively studied permafrost and applied the results of their learning to the development of Russia's north.

In a similar way, prospectors and explorers were aware of permafrost in the northern regions of North America for many years, but it was not until after World War II that systematic studies of perennially frozen ground were undertaken by scientists and engineers in the United States and Canada. Since exploitation of the great petroleum resources on the northern continental shelves began in earnest in the 1970s, investigations into subsea permafrost have progressed even more rapidly than have studies of permafrost on land.

Alpine permafrost studies had their beginning in the study of rock glaciers in the Alps of Switzerland. Although ice was known to exist in rock glaciers, it was not until after World War II that investigation by geophysical methods clearly demonstrated slow movement of perennial ice—*i.e.*, permafrost. In the 1970s and '80s, detailed geophysical work and temperature and borehole examination of mountain permafrost began in Russia, China, and Scandinavia, especially with regard to construction in high mountain and plateau areas.

ORIGIN AND STABILITY OF PERMAFROST

Air temperature and ground temperature. In areas where the mean annual air temperature becomes colder than 0° C, some of the ground frozen in the winter will not be completely thawed in the summer; therefore, a layer of permafrost will form and continue to grow downward gradually each year from the seasonally frozen ground. The permafrost layer will become thicker each winter, its thickness controlled by the thermal balance between the heat flow from the Earth's interior and that flowing outward into the atmosphere. This balance depends on the mean annual air temperature and the geothermal gradient. The average geothermal gradient is an increase of 1° C (1.8° F) for every 30 to 60 metres of depth. Eventually the thickening permafrost layer reaches an equilibrium depth at which the amount of geothermal heat reaching the permafrost is on the average equal to that lost to the atmosphere. Thousands of years are required to attain a state of equilibrium where permafrost is hundreds of feet thick.

Geothermal gradient

An example of the change of temperature of frozen ground with depth and the upper and lower limit of permafrost is illustrated in Figure 16. The annual fluctuation of air temperature from winter to summer is reflected in a subdued manner in the upper few metres of the ground. This fluctuation diminishes rapidly with depth, being only a few degrees at 7.5 metres, and is barely detectable at 15 metres. The level of zero amplitude, at which fluctuations are hardly detectable, is 9 to 15 metres. If the permafrost is in thermal equilibrium, the temperature at the level of zero amplitude is generally regarded as the minimum temperature of the permafrost. Below this depth the temperature increases steadily under the influence of heat from the Earth's interior. The temperature of permafrost at the depth of minimum annual seasonal change varies from near 0° C at the southern limit of permafrost to −10° C (14° F) in northern Alaska and −13° C (9° F) in northeastern Siberia.

As the climate becomes colder or warmer, but maintaining a mean annual temperature colder than 0° C, the temperature of the permafrost correspondingly rises or declines, resulting in changes in the position of the base of permafrost. The position of the top of permafrost will be lowered by thawing when the climate warms to a mean annual air temperature warmer than 0° C. The rate at which the base or top of permafrost is changed depends not only on the amount of climatic fluctuation but also on the amount of ice in the ground and the composition of the ground, conditions that in part control the geothermal gradient. If the geothermal gradient is known and if the surface temperature remains stable for a long period of time, it is, therefore, possible to predict from a knowledge of the mean annual air temperature the thickness of permafrost in a particular area that is remote from bodies of water.

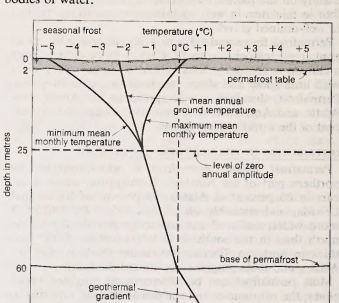


Figure 16: Hypothetical example of a temperature profile and thickness of permafrost in central Alaska.

Climatic change. Permafrost is the result of present climate. Many temperature profiles show, however, that permafrost is not in equilibrium with present climate at the sites of measurement. Some areas show, for example, that climatic warming since the last third of the 19th century has caused a warming of the permafrost to a depth of more than 100 metres (see Figure 17). In such areas much of the permafrost is a product of a colder, former climate.

Origin of subsea permafrost

The distribution and characteristics of subsea permafrost point to a similar origin. At the height of the glacial epoch, especially about 20,000 years ago, most of the continental shelf in the Arctic Ocean was exposed to polar climates for thousands of years. These climates caused cold permafrost to form to depths of more than 700 metres. Subsequently, within the past 10,000 years, the Arctic Ocean rose and advanced over a frozen landscape to produce a degrading relict subsea permafrost. The perennially frozen ground is no longer exposed to a cold atmosphere, and the salt water has caused a reduction in strength and consequent melting of the ice-rich permafrost (which is bonded by freshwater ice). The temperature of subsea permafrost, near −1° C (30° F), is no longer as low as it was in glacial times and is therefore sensitive to warming from geothermal heat and to the encroaching activities of humans.

It is thought that permafrost first occurred in conjunction with the onset of glacial conditions about three million years ago, during the late Pliocene Epoch. In the subarctic at least, most permafrost probably disappeared during interglacial times and reappeared in glacial times. Most existing permafrost in the subarctic probably formed in the cold (glacial) period of the past 100,000 years.

LOCAL THICKNESS

The thickness and areal distribution of permafrost are directly affected by snow and vegetation cover, topography, bodies of water, the interior heat of the Earth, and the temperature of the atmosphere, as mentioned earlier.

Effects of climate. The most conspicuous change in thickness of permafrost is related to climate. At Barrow, Alaska, U.S., the mean annual air temperature is −12° C (10° F), and the thickness is 400 metres. At Fairbanks, Alaska, in the discontinuous zone of permafrost in central Alaska, the mean annual air temperature is −3° C (27° F), and the thickness is about 90 metres. Near the southern border of permafrost, the mean annual air temperature is about 0° or −1° C, and the perennially frozen ground is only a few feet thick.

If the mean annual air temperature is the same in two areas, the permafrost will be thicker where the conductivity of the ground is higher and the geothermal gradient is less. A.H. Lachenbruch of the U.S. Geological Survey reports an interesting example from northern Alaska (see Figure 17). The mean annual air temperatures at Cape Simpson and Prudhoe Bay are similar, but permafrost thickness is 275 metres at Cape Simpson and about 650 metres at Prudhoe Bay because rocks at Prudhoe Bay are more siliceous and have a higher conductivity and a lower geothermal gradient than rocks at Cape Simpson.

Soil under deep lakes

Effects of water bodies. Bodies of water, lakes, rivers, and the sea have a profound effect on the distribution of permafrost. A deep lake that does not freeze to the bottom during the winter will be underlain by a zone of thawed material. If the minimum horizontal dimension of the deep lake is about twice as much as the thickness of permafrost nearby, there probably exists an unfrozen vertical zone extending all the way to the bottom of permafrost. Such thawed areas extending all the way through permafrost are widespread under rivers and sites of recent rivers in the discontinuous zone of permafrost and under major, deep rivers in the far north. Under the wide floodplains of rivers in the subarctic, the permafrost is sporadically distributed both laterally and vertically. Small, shallow lakes that freeze to the bottom each winter are underlain by a zone of thawed material, but the thawed zone does not completely penetrate permafrost except near the southern border of permafrost.

Effects of solar radiation, vegetation, and snow cover. Inasmuch as south-facing hillslopes receive more incoming solar energy per unit area than other slopes, they are warmer; permafrost is generally absent on these in the discontinuous zone (Figure 18) and is thinner in the continuous zone. The main role of vegetation in permafrost areas is to shield perennially frozen ground from solar energy. Vegetation is an excellent insulating medium and removal or disturbance of it, either by natural processes or by humans, causes thawing of the underlying permafrost. In the continuous zone the permafrost table may merely be lowered by the disturbance of vegetation, but in a discontinuous zone permafrost may be completely destroyed in certain areas.

Snow cover also influences heat flow between the ground and the atmosphere and therefore affects the distribution of permafrost. If the net effect of timely snowfalls is to prevent heat from leaving the ground in the cold winter, permafrost becomes warmer. Actually, local differences in vegetation and snowfall in areas of thin and warm permafrost are critical for the formation and existence of the perennially frozen ground. Permafrost is not present in areas of the world where great snow thicknesses persist throughout most of the winter.

ICE CONTENT

Types of ground ice. The ice content of permafrost is probably the most important feature of permafrost affecting human life in the north. Ice in the perennially frozen ground exists in various sizes and shapes and has definite distribution characteristics. The forms of ground ice can

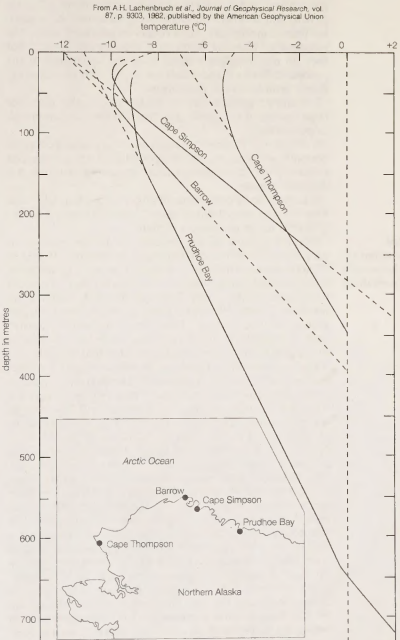


Figure 17: Generalized temperature profiles through permafrost at four sites on the Alaskan Arctic coast. Solid lines represent measured temperatures; dashed lines represent extrapolations.

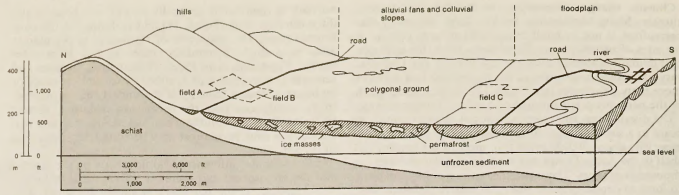


Figure 18: Character and distribution of permafrost in the Fairbanks, Alaska, area.
From T.L. Pewe, *Effect of Permafrost on Cultivated Fields, Fairbanks, Alaska* (1954); U.S. Geological Survey

be grouped into five main types: (1) pore ice, (2) segregated, or Taber, ice, (3) foliated, or wedge, ice, (4) pingo ice, and (5) buried ice.

1. Pore ice, which fills or partially fills pore spaces in the ground, is formed by pore water freezing in situ with no addition of water. The ground contains no more water in the solid state than it could hold in the liquid state.

2. Segregated, or Taber, ice includes ice films, seams, lenses, pods, or layers generally 0.15 to 13 centimetres (0.06 to 5 inches) thick that grow in the ground by drawing in water as the ground freezes. Small ice segregations are the least spectacular but one of the most extensive types of ground ice, and engineers and geologists interested in ice growth and its effect on engineering structures have studied them considerably. Such observers generally accept the principle of bringing water to a growing ice crystal, but they do not completely agree as to the mechanics of the processes. Pore ice and Taber ice occur both in seasonally frozen ground and in permafrost.

3. Foliated ground ice, or wedge ice, is the term for large masses of ice growing in thermal contraction cracks in permafrost.

4. Pingo ice is clear, or relatively clear, and occurs in permafrost more or less horizontally or in lens-shaped masses. Such ice originates from groundwater under hydrostatic pressure.

5. Buried ice in permafrost includes buried sea, lake, and river ice and recrystallized snow, as well as buried blocks of glacier ice in permafrost climate.

Ice content of the world's permafrost

World estimates of the amount of ice in permafrost vary from 200,000 to 500,000 cubic kilometres (49,000 to 122,000 cubic miles), or less than 1 percent of the total volume of the Earth. It has been estimated that 10 percent by volume of the upper 3 metres of permafrost on the northern Coastal Plain of Alaska is composed of foliated ground ice (ice wedges). Taber ice is the most extensive type of ground ice, and in places it represents 75 percent of the ground by volume. It is calculated that the pore and Taber ice content in the depth between 0.5 and 3 metres (surface to 0.5 metre is seasonally thawed) is 61 percent by volume, and between 3 and 9 metres it is 41 percent. The total amount of pingo ice is less than 0.1 percent of the permafrost. The total ice content in the permafrost of the Arctic Coastal Plain of Alaska is estimated to be 1,500 cubic kilometres, and below 9 metres most of that is present as pore ice.

Ice wedges. The most conspicuous and controversial type of ground ice in permafrost is that formed in large ice wedges or masses with parallel or subparallel foliation structures. Most foliated ice masses occur as wedge-shaped, vertical, or inclined sheets or dikes 2.5 centimetres to 3 metres wide and 0.3 to 9 metres high when viewed in transverse cross section (Figure 19). Some masses seen on the face of frozen cliffs may appear as horizontal bodies a few centimetres to 3 metres in thickness and 0.3 to 14 metres long, but the true shape of these ice wedges can be seen only in three dimensions. Ice wedges are parts of polygonal networks of ice enclosing cells of frozen ground 3 to 30 metres or more in diameter.

Origins. The origin of ground ice was first studied in Siberia, and discussions in print of the origin of large ground-ice masses in perennially frozen ground of North America have gone on since Otto von Kotzebue recorded ground ice in 1816 at a spot now called Elephant's Point in Eschscholtz Bay of Seward Peninsula. The theory for the origin of ice wedges now generally accepted is the thermal contraction theory that, during the cold winter, polygonal thermal contraction cracks, a centimetre or two wide and a few metres deep, form in the frozen ground; then when, in early spring, water from the melting snow runs down these tension cracks and freezes, a vertical vein of ice is produced that penetrates into permafrost; when the permafrost warms and re-expands during the following summer, horizontal compression produces upturning of the frozen sediment by plastic deformation; then during the next winter, renewed thermal tension reopens the vertical ice-cemented crack, which may be a zone of weakness; another increment of ice is added in the spring when meltwater again enters and freezes. Over the years the vertical wedge-shaped mass of ice is produced (Figure 19).

Thermal contraction theory

Active wedges, inactive wedges, and ice-wedge casts. Ice wedges may be classified as active, inactive, and ice-wedge casts. Active ice wedges are those that are actively growing. The wedge may not crack every year, but during many or most years cracking does occur, and an increment of ice is added. Ice wedges require a much more rigorous climate

Troy L. Pewe, Arizona State University



Figure 19: Foliated ground-ice mass (ice wedge) in rich organic silt exposed by gold-mining operations on Wilbur Creek, near Livengood, Alaska.

to grow than does permafrost. The permafrost table must be chilled to −15° to −20° C (5° to −4° F) for contraction cracks to form. On the average, it is assumed that ice wedges generally grow in a climate where the mean annual air temperature is −6° or −8° C (21° or 18° F) or colder. In regions with a general mean annual temperature only slightly warmer than −6° C, ice wedges occasionally form in restricted cold microclimate areas or during cold periods of a few years' duration.

The area of active ice wedges appears to roughly coincide with the continuous permafrost zone. From north to south across the permafrost area in North America, a decreasing number of wedges crack frequently. The line dividing zones of active and inactive ice wedges is arbitrarily placed at the position where it is thought most wedges do not frequently crack.

Inactive ice wedges are those that are no longer growing. The wedge does not crack in winter and, therefore, no new ice is added. A gradation between active ice wedges and inactive ice wedges occurs in those wedges that crack rarely. Inactive ice wedges have no ice seam or crack extending from the wedge upward to the surface in the spring. The wedge top may be flat (Figure 19), especially if thawing has lowered the upper surface of the wedge at some time in the past.

Ice wedges in the world are of several ages, but none appear older than the onset of the last major cold period, about 100,000 years ago. Wedges dated by radiocarbon analyses range from 3,000 to 32,000 years in age.

Indication of a cooler past

In many places in the now temperate latitudes of the world, in areas of past permafrost, ice wedges have melted, and resulting voids have been filled with sediments collapsing from above and the sides. These ice-wedge casts are important as paleoclimatic indicators and indicate a climate of the past with at least a mean annual air temperature of −6° or −8° C or colder.

SURFACE MANIFESTATIONS OF PERMAFROST AND SEASONALLY FROZEN GROUND

Many distinctive surface manifestations of permafrost exist in the Arctic and subarctic, including such geomorphic features as polygonal ground, thermokarst phenomena, and pingos. In addition to the above, there are many features caused in large part by frost action that are common in but not restricted to permafrost areas, such as solifluction (soil flowage) and frost-sorted patterned ground.

Areas underlain by permafrost. *Polygonal ground.* One of the most widespread geomorphic features associated with permafrost is the microrelief pattern on the surface of the ground generally called polygonal ground, or tundra polygons (Figure 20). This pattern, which covers thousands of square miles of the Arctic and less in the subarctic, is caused by an intersecting network of shallow troughs delineating polygons 3 to 30 metres in diameter. The troughs are underlain by more or less vertical ice wedges (Figure 19) 0.6 to 3 metres across on the top that are joined together in a honeycomb network. These large-scale polygons should not be confused with the small-scale polygons or patterned ground produced by frost sorting.

Ice-wedge polygons

The ice-wedge polygons may be low-centred or high-centred. Upturning of strata adjacent to the ice wedge may make a ridge of ground on the surface on each side of the wedge, thus enclosing the polygons. Such polygons are lower in the centre and are called low-centre polygons or raised-edge polygons and may contain a pond in the centre. Low-centre, or raised-edge, polygons indicate that ice wedges are actually growing and that the sediments are being actively upturned. If erosion, deposition, or thawing is more prevalent than the up-pushing of the sediments along the side of the wedge or if the material being pushed up cannot maintain itself in a low ridge, the low ridges will be absent, and there may be either no polygons at the surface or the polygons may be higher in the centre than the troughs over the ice wedges that enclose them. Both high-centre and low-centre tundra polygons are widespread in the polar areas and are good indicators of the presence of foliated ice masses; care must be taken, however, to demonstrate that the pattern is not a relic and an indication of ice-wedge casts.

In many parts of the temperate latitudes of Asia, Europe, and North America, incompletely developed or poorly developed polygonal ground occurs on the same scale as in the Arctic. These large-scale polygons in the nonpermafrost areas are excellent evidence of the former extent of permafrost and ice wedges in the past glacial period.

In many areas of the continuous permafrost zone surface drainage follows the troughs of the polygons (tops of the ice wedges); and at ice wedge junctions, or elsewhere, melting may occur to form small pools. The joining of these small pools by a stream causes the pools to resemble beads on a string, a type of stream form called beaded drainage. Such drainage indicates the presence of perennially frozen, fine-grained sediments cut by ice wedges.

Thermokarst formations. The thawing of permafrost creates thermokarst topography, an uneven surface that contains mounds, sinkholes, tunnels, caverns, and steep-walled ravines caused by melting of ground ice. The hummocky ground surface resembles karst topography in limestone areas. Thawing may result from artificial or natural removal of vegetation or from a warming climate.

Depressions and pits

Thawed depressions filled with water (thaw lakes, thermokarst lakes, cave-in lakes) are widespread in permafrost areas, especially in those underlain with perennially frozen silt. They may occur on hillsides or even on hilltops and are good indicators of ice-rich permafrost. Locally, deep thermokarst pits 6 metres deep and 9 metres across may form as ground ice melts. These openings may exist as undetected caverns for many years before the roof collapses. Such collapses in agricultural or construction areas are real dangers. Thermokarst mounds are polygonal or circular hummocks 3 to 15 metres in diameter and 0.3 to 2.5 metres high that are formed as a polygonal network of ice melts and leaves the inner-ice areas as mounds.

Pingos. The most spectacular landforms associated with permafrost are pingos, small ice-cored circular or elliptical hills of frozen sediments or even bedrock, 3 to more than 60 metres high and 15 to 450 metres in diameter. Pingos are widespread in the continuous permafrost zone and are quite conspicuous because they rise above the tundra. They are much less conspicuous in the forested area of the discontinuous permafrost zone. They are generally cracked on top with summit craters formed by melting ice. There are two types of pingos, based on origin. The closed-system type forms in level areas when unfrozen groundwater in a thawed zone becomes confined on all sides by permafrost, freezes, and heaves the frozen overburden to form a mound. This type is larger and occurs mainly in tundra areas of continuous permafrost. The open-system type is generally smaller and forms on slopes when water beneath or within the permafrost penetrates

By courtesy of the U.S. Army (CRREL); photograph, Robert Lewellen



Figure 20: Raised-edge ice wedge polygons on the seacoast near Barrow, Alaska, in summer. The polygons are from 7 to 15 metres in diameter.

the permafrost under hydrostatic pressure. A hydrolaccolith (water mound) forms and freezes, heaving the overlying frozen and unfrozen ground to produce a mound.

Present pingos are apparently the result of postglacial climate and are less than 4,000–7,000 years old. Pingos were present in now temperate latitudes during the latest glacial epoch and are now represented by low circular ridges enclosing bogs or lowlands.

Near the southern border of permafrost occur palsas, low hills and knobs of perennially frozen peat about 1.5 to 6 metres high, evidently forming with accumulation of peat and segregation of ice.

Features related to seasonal frost. Many microgeomorphic features common to the periglacial environment may or may not be associated with permafrost.

Patterned ground. Intense seasonal frost action, repeated freezing and thawing throughout the year, produces small-scale patterned ground. Repetitive freezing and thawing tends to stir and sort granular sediments, thus forming circles, stone nets, and polygons a few centimetres to 6 metres in diameter. The coarse cobbles and boulders form the outside of the ring and the finer sediments occur in the centre. The features require a rigorous climate with some fine-grained sediments and soil moisture, but they do not necessarily need underlying permafrost. Permafrost, however, forms an impermeable substratum that keeps the soil moisture available for frost action. On gentle slopes the stone nets may be distorted into garlands by downslope movement or, if the slope is steep, into stone stripes about half a metre wide and 30 metres long.

Soil flow. In areas underlain by an impermeable layer (seasonally frozen ground or perennially frozen ground), the active layer is often saturated with moisture and is quite mobile. The progressive downslope movement of saturated detrital material under the action of gravity and working in conjunction with frost action is called solifluction. This material moves in a semifluid condition and is manifested by lobelike and sheetlike flows of soil on slopes. The lobes are up to 30 metres wide and have a steep front 0.3 to 1.5 metres high. An outstanding feature of solifluction is the mass transport of material over low-angle slopes. Solifluction deposits are widespread in polar areas and consist of a blanket 0.3 to 1.8 metres thick of unstratified or poorly stratified, unsorted, heterogeneous, till-like detrital material of local origin. In many areas the terrain is characterized by relatively smooth, round hills and slopes with well-defined to poorly defined solifluction lobes or terraces. If the debris is blocky and angular and fine material is absent, the lobes are poorly developed or absent. Areas in which solifluction lobes are well formed lie almost entirely above or beyond the forest limit.

Solifluction lobes

In many areas the frost-rived debris contains few fine materials and little water and consists of angular fragments of well-jointed, resistant rock. Under such circumstances, solifluction lobes do not often occur, but instead striking sheets or streams of angular rubble form. These are called rock streams or rubble sheets.

PROBLEMS POSED BY PERMAFROST

Permafrost engineering. *General issues.* Development of the north demands an understanding of and the ability to cope with problems of the environment dictated by permafrost. Although the frozen ground hinders agricultural and mining activities, the most dramatic, widespread, and economically important examples of the influence of permafrost on life in the north involve construction and maintenance of roads, railroads, airfields, bridges, buildings, dams, sewers, and communication lines. Engineering problems are of four fundamental types: (1) those involving thawing of ice-rich permafrost and subsequent subsidence of the surface under unheated structures such as roads and airfields, (2) those involving subsidence under heated structures, (3) those resulting from frost action, generally intensified by poor drainage caused by permafrost, and (4) those involved only with the temperature of permafrost that causes buried sewer, water, and oil lines to freeze.

A thorough study of the frozen ground should be part of the planning of any engineering project in the north. It is generally best to attempt to disturb the permafrost as little as possible in order to maintain a stable foundation for engineering structures, unless the permafrost is thin; then, it may be possible to destroy the permafrost. The method of construction preserving the permafrost has been called the passive method; alternately, the destroying of permafrost is the active method.

Permafrost thawing and frost heaving. Because thawing of permafrost and frost action are involved in almost all engineering problems in polar areas, it is advisable to consider these phenomena generally. The delicate thermal equilibrium of permafrost is disrupted when the vegetation, snow cover, or active layer is compacted. The permafrost table is lowered, the active layer is thickened, and considerable ice is melted. This process lowers the surface and provides (in summer) a wetter active layer with less bearing strength. Such disturbance permits a greater penetration of summer warming. It is common procedure to place a fill, or pad, of gravel under engineering works. Such a fill generally is a good conductor of heat and, if thin, may cause additional thawing of permafrost. The fill must be made thick enough to contain the entire amplitude of seasonal temperature variation—in other words, thick enough to restrict the annual seasonal freezing and thawing to the fill and the compacted active layer. Under these conditions no permafrost will thaw. Such a procedure is quite feasible in the Arctic, but in the warmer subarctic it is impractical because of the enormous amounts of fill needed. Under a heated building, profound thawing may occur more rapidly than under roads and airfields.

Insulation with gravel pads

Frost action, the freezing and thawing of moisture in the ground, has long been known to seriously disrupt and destroy structures in both polar and temperate latitudes. In the winter the freezing of ground moisture produces upward displacement of the ground (frost heaving), and in the summer excessive moisture in the ground brought in during the freezing operation causes loss of bearing strength. Frost action is best developed in silt-sized and silty clay-sized sediments in areas of rigorous climate and poor drainage. Polar latitudes are ideal for maximum frost action because most lowland areas are covered by fine-grained sediments, and the underlying permafrost causes poor drainage.

Development in permafrost areas. *Structures on piles.* Piles are used to support many, if not most, structures built on ice-rich permafrost. In regions of cold winters, many pile foundations are in ground subject to seasonal freezing and, therefore, possibly subject to the damaging effect of frost heaving, which tends to displace the pile upward and thus to disturb the foundation of the structure. The displacement of piling is not limited to the far north, though maximum disturbance probably is encountered most widely in the subarctic. Expensive maintenance and sometimes complete destruction of bridges, school buildings, military installations, pipelines, and other structures have resulted from failure to understand the principles of frost heaving of piling.

A remarkable construction achievement in a permafrost environment is the Trans-Alaska Pipeline System. Completed in 1977, this 1,285-kilometre-long, 122-centimetre-diameter pipeline transports crude oil from Prudhoe Bay to an ice-free port at Valdez. The pipeline was originally designed for burial along most of the route. However, because the oil is transported at 70° to 80° C (158° to 176° F), such an installation would have thawed the adjacent permafrost, causing liquefaction, loss of bearing strength, and soil flow. To prevent destruction of the pipeline, about half of the line (615 kilometres) is elevated onto beams held up by vertical support members. The pipeline safely discharges its heat into the air, while frost heaving of the 120,000 vertical support members is prevented by freezing them firmly into the permafrost through the use of special heat-radiating thermal devices.

The Trans-Alaska Pipeline System

Highways and railroads. Highways in polar areas are relatively few and mainly unpaved. They are subject to subsidence by thawing of permafrost in summer, frost heaving in winter, and loss of bearing strength on fine-grained sediments in summer. Constant grading of gravel roads permits maintenance of a relatively smooth highway. Where the road is paved over ice-rich permafrost,

the roadway becomes rough and is much more costly to maintain than are unpaved roads. Many of the paved roads in polar areas have required resurfacing two or three times in a 10-year period.

Railroads particularly have serious construction problems and require costly upkeep in permafrost areas because of the necessity of maintaining a relatively low gradient and the subsequent location of the roadbed in ice-rich lowlands that are underlain with perennially frozen ground. The Trans-Siberian Railroad, the Alaska Railroad, and some Canadian railroads in the north are locally underlain by permafrost with considerable ground ice. As the large masses of ice melt each summer, constant maintenance is required to level these tracks. In winter, extensive maintenance is also required to combat frost heaving when local displacements of 2.5 to 35 centimetres occur in roadbeds and bridges.

Agriculture. Permafrost affects agricultural developments in many parts of the discontinuous permafrost zone. Its destructive effect on cultivated fields in both Russia and North America results from the thawing of large masses of ice in the permafrost. If care is not exercised in selecting areas to be cleared for cultivation, thawing of the permafrost may necessitate abandonment of fields or their reduction to pasturage. As illustrated in Figure 18, field A, on the southward-facing hillside, is not affected by permafrost; field B, on the alluvial fan, however, is underlain with permafrost containing large masses of ground ice. When field B is cleared, thermokarst topography will form. Field C, on the floodplain, is underlain with permafrost without large ice masses, and so no thermokarst pits and mounds will form when this ground is cleared.

Offshore structures. One of the most active and exciting areas of permafrost engineering is in subsea permafrost. Knowledge of the distribution, type, and water or ice content of subsea permafrost is critical for planning petroleum exploration, locating production structures, burying pipelines, and driving tunnels beneath the seabed. Furthermore, the temperature of the seabed must be known in order to predict potential sites of accumulation of gas hydrates or areas in which groundwater or artesian pressures are likely. In addition, knowledge of the distribution of subsea permafrost permits a thorough interpretation of regional geologic history. (T.L.Pe.)

BIBLIOGRAPHY

General works. SAMUEL C. COLBECK (ed.), *Dynamics of Snow and Ice Masses* (1980), includes chapters on valley glaciers, ice sheets, snow packs, icebergs, sea ice, and avalanches, emphasizing the basic physics. P.V. HOBBS, *Ice Physics* (1974), treats all aspects of the physics and chemistry of ice. W. RICHARD PELTIER (ed.), *Ice in the Climate System* (1993), is a modern review of the past, present, and future interactions between ice and climate.

Ice in lakes and rivers. E.R. POUNDER, *The Physics of Ice* (1965), treats concisely the structure and physical properties of ice. GEORGE D. ASHTON (ed.), *River and Lake Ice Engineering* (1986), provides a comprehensive treatment of the general principles for engineering applied to river and lake ice problems. BERNARD MICHEL, *Winter Regime of Rivers and Lakes* (1971), treats freshwater ice. The essay by GEORGE D. ASHTON, "Freshwater Ice Growth and Decay," in the work by Colbeck (cited above), summarizes the general principles of river ice behaviour, including ice accumulation processes and thermal effects. For information regarding the geographic distribution of lake and river ice, the following works are useful: W.T.R. ALLEN, *Freeze-up, Break-up, and Ice Thickness in Canada* (1977?); and, for the former Soviet Union, I.P. GERASIMOV *et al.* (eds.), *Физико-географический Атлас Мира* (1964); and L.N. MESIATSEVA, *Атлас СССР* (1984). (G.D.A.)

Glaciers and ice sheets. W.S.B. PATERSON, *The Physics of Glaciers*, 2nd ed. (1981), is the standard text on glaciers and ice sheets, emphasizing process rather than description. J.T. ANDREWS, *Glacial Systems* (1975), is a compact, clear introduction to glaciers and their environment. MICHAEL HAMBREY and JÜRG ALEAN, *Glaciers* (1992); and ROBERT P. SHARP, *Living Ice* (1988), are well-illustrated introductions to glaciers and how they affect the landscape. NATIONAL RESEARCH COUNCIL (U.S.), AD HOC COMMITTEE ON THE RELATIONSHIP BETWEEN LAND ICE AND SEA LEVEL, *Glaciers, Ice Sheets, and Sea Level: Effect of a CO_2-Induced Climatic Change* (1985); and J.T. HOUGHTON, G.J. JENKINS, and J.J. EPHRAUMS (eds.), *Climate Change: The IPCC Scientific Assessment* (1990), discuss the interaction between climate, glaciers, and sea level. D.J. DREWRY (ed.), *Antarctica: Glaciological and Geophysical Folio* (1983), is a large-format compendium on the Earth's largest ice mass. "Fast Glacier Flow: Ice Streams, Surging, and Tidewater Glaciers," *Journal of Geophysical Research*, part B, *Solid Earth and Planets*, 92(9):8835–8841 (1987), is a collection of review papers and scientific contributions from the Chapman Conference on Fast Glacier Flow. (M.F.M.)

Icebergs and pack ice. The published literature on icebergs is found mainly in U.S. journals and in atlases prepared by the U.S. Coast Guard and U.S. Navy for Arctic and Antarctic military and scientific maritime resupply expeditions. A classic discussion of icebergs and sea ice in the Arctic Ocean is *The "Marion" Expedition to Davis Strait and Baffin Bay*, vol. 3 by E.H. SMITH, *Arctic Ice* (1931). Also of interest is the treatment of the oceanography of the north polar basin by FRIDTJOF NANSEN, *The Norwegian North Polar Expedition, 1893–1896: Scientific Results*, vol. 3 (1902, reprinted 1969), and *Farthest North*, 2 vol. (1897, reissued 1967). The importance of ice in the water budget of the planet may be studied by consulting JAMES L. DYSON, *The World of Ice* (1962, reissued 1972). ROBERT C. PRITCHARD (ed.), *Sea Ice Processes and Models* (1980), a collection of proceedings papers, reports field observations and the development of models in an effort to establish usable flow laws for sea ice. Two reference works are UNITED STATES HYDROGRAPHIC OFFICE, *A Functional Glossary of Ice Terminology* (1952); and WORLD METEOROLOGICAL ORGANIZATION, *Sea-Ice Nomenclature* (1970).

The age of icebergs can be understood by consulting W. DANSGAARD *et al.*, "One Thousand Centuries of Climatic Record from Camp Century on Greenland Ice Sheet," *Science*, 166(3903):377–381 (1969); and P.F. SCHOLANDER *et al.*, "Composition of Gas Bubbles in Greenland Icebergs," *The Journal of Glaciology*, 3:813–822 (1961). A major reference work on the proposed use of icebergs as a freshwater source is the collected conference papers in A.A. HUSSEINY (ed.), *Iceberg Utilization* (1978). (T.F.B./Ed.)

Permafrost. A.L. WASHBURN, *Geocryology* (1979), is the most thorough book in English on permafrost and periglacial processes. H.M. FRENCH, *The Periglacial Environment* (1976), clearly summarizes permafrost and periglacial processes, with emphasis on examples from Canada. The greatest source of permafrost information is the proceedings of the various International Conference on Permafrost meetings; each volume contains numerous up-to-date papers in English from many different countries. WILFRIED HAEBERLI, *Creep of Mountain Permafrost: Internal Structure and Flow of Alpine Rock Glaciers* (1985), is a discussion of rock glaciers, a prominent feature of Alpine permafrost. TROY L. PÉWÉ, "Alpine Permafrost in the Contiguous United States: A Review," *Arctic and Alpine Research*, 15(2):145–156 (1983), summarizes in detail the character and distribution of mountain permafrost in this region.

ARTHUR H. LACHENBRUCH, *Mechanics of Thermal Contraction Cracks and Ice-Wedge Polygons in Permafrost* (1962), is a classic paper on the quantitative interpretation of the formation of ice-wedge polygons in permafrost. TROY L. PÉWÉ, RICHARD E. CHURCH, and MARVIN J. ANDRESEN, *Origin and Paleoclimatic Significance of Large-Scale Patterned Ground in the Donnelly Dome Area, Alaska* (1969), discusses the origin of ice-wedge casts and relict permafrost in central Alaska and offers paleoclimatic interpretations. R. DALE GUTHRIE, *Frozen Fauna of the Mammoth Steppe* (1990), discusses fossil carcasses of Ice Age mammals preserved in permafrost. TROY L. PÉWÉ, *Geologic Hazards of the Fairbanks Area, Alaska* (1982), a highly illustrated work, contains an up-to-date presentation of the greatest geologic hazard to life in polar areas: problems posed by seasonally and perennially frozen ground. G.H. JOHNSTON (ed.), *Permafrost: Engineering Design and Construction* (1981), is a comprehensive book on construction problems in permafrost areas, with examples mainly from northern Canada. TROY L. PÉWÉ, "Permafrost," in GEORGE A. KIERSCH *et al.* (eds.), *The Heritage of Engineering Geology: The First Hundred Years* (1991), pp. 277–298, provides an up-to-date, well-illustrated treatment of the origin, distribution, and ice content of permafrost and of engineering problems in permafrost regions. (T.L.Pe.)

Iceland

Iceland (Icelandic: Ísland), an island country located in the North Atlantic Ocean, is a land of vivid contrasts. Sparkling glaciers lie across its ruggedly beautiful mountain ranges, while a vast quantity of subterranean thermal activity makes Iceland one of the most active volcanic regions in the world. Glaciers and cooled lava beds each cover approximately one-tenth of the 39,768 square miles (103,000 square kilometres) of the country. The glaciers are a reminder of Iceland's close proximity to the Arctic Circle, while the volcanoes, reaching deep into the unstable interior of the Earth, are explained by the fact that Iceland is located on the top of the Mid-Atlantic Ridge. It is estimated that since the year 1500 about one-third of the Earth's total lava flow has poured out of the volcanoes of Iceland.

Iceland was founded more than 1,000 years ago during the Viking Age of exploration. The capital, Reykjavík ("Bay of Smokes"), is near the site of the island's first farmstead. The early settlement, made up primarily of Norwegian seafarers and adventurers, fostered further excursions to Greenland and the coasts of North America, or Vinland. In spite of its physical isolation some 500 miles (800 kilometres) from Scotland, its nearest European neighbour, Iceland has remained throughout its history very much a part of European civilization. It is a Scandinavian country, modern in nearly every respect.

This article is divided into the following sections:

Physical and human geography 760
- The land 760
  - Relief
  - Drainage and soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 762
- The economy 763
  - Resources
  - Fishing
  - Agriculture
  - Industry
  - Finance
  - Transportation and tourism
  - Foreign trade
- Administration and social conditions 764
  - Government
  - Foreign relations
  - Health and welfare
  - Education
- Cultural life 764
  - The arts
  - Sports
  - Cultural institutions

History 765
- Early history 765
  - Settlement (*c.* 870–*c.* 930)
  - Commonwealth (*c.* 930–1264)
- Iceland under foreign rule 765
  - Late Middle Ages (1264–*c.* 1550)
  - Growth of Danish royal power (*c.* 1550–*c.* 1830)
- Modern Iceland 766
  - Struggle for independence (*c.* 1830–1904)
  - Home rule and sovereignty (1904–44)
  - The Icelandic republic

Bibliography 767

Physical and human geography

THE LAND

Geologically young, Iceland contains about 200 volcanoes of various types. A new volcano erupting on the bottom of the sea between November 1963 and June 1967 created the island of Surtsey, off the southwestern coast. The new island grew to nearly one square mile in area and rose more than 560 feet (170 metres) above sea level, a total of 950 feet from the ocean floor.

Volcanism

Volcanic activity has been particularly frequent since the 1970s. A major eruption took place in 1973, when a volcano on Heima Island (Heimaey) spilled lava into the town of Vestmannaeyjar, an important fishing centre. Most of the 5,300 residents had to be evacuated, and—although the harbour remained intact—about one-third of the town was destroyed. Continuous eruptions took place in the Krafla area in the northeast in 1975–84. Iceland's best-known volcano, Hekla, erupted four times in the 20th century: in 1947, 1970, 1980, and 1991.

Relief. Iceland is largely a tableland broken up by structural faults. Its average elevation is 1,640 feet above sea level, but one-fourth of the country lies below 650 feet. The highest point is 6,952 feet (2,119 metres), at Hvannadals Peak, the top of Öræfajökull in Vatna Glacier. The glaciers range in size from small ones in mountain recesses to enormous glacial caps topping extensive mountain ranges. Vatna Glacier covers an area of about 3,200 square miles (8,288 square kilometres) and is about 3,000 feet (914 metres) deep at its thickest point.

Much of Iceland is underlain by basalt, a dark rock of igneous origin. The oldest rocks were formed about 16 million years ago. The landscape in basaltic areas is one of plateau and fjord, characterized by successive layers of lava visible one above the other on the valley sides. The basalt sheets tend to tilt somewhat toward the centre of the country. The U shape of the valleys is largely the result of glacial erosion. The depressed zones between the basalt areas have extensive plateaus above which rise single volcanoes, table mountains, or other mountain masses with steep sides.

Iceland has more hot springs and solfataras—volcanic vents that emit hot gases and vapours—than any other country. Alkaline hot springs are found in some 250 areas throughout the country. The largest, Deildartunguhver, emits about 48 gallons (180 litres) of boiling water per second. The total power output of the Torfa Glacier area, the largest of the 19 high-temperature solfatara regions, is estimated to equal about 1,000 megawatts. Earthquakes are frequent in Iceland but rarely result in serious damage. Most of the buildings erected since the mid-20th century have been built of reinforced concrete and designed to withstand severe shocks from earthquakes.

Traditional regions

Traditionally, Iceland has been divided according to the four points of the compass. The centre of the country is uninhabited. In the southwest several fine natural harbours have directed interest toward the sea, and good fishing grounds lie off the shores of this region. Because of its extensive lava fields and heaths, the southwest has little farmland. The middle west is divided between fishing and farming and has many places of great natural beauty. The western fjords have numerous well-sheltered harbours and good fishing grounds but little lowland suitable for agriculture. The north is divided into several smaller districts, each of which has relatively good farmland. The eastern fjords resemble the western fjords but have, in addition, an inner lowland. The southeast, locked between the glaciers and the sea, has a landscape of rugged splendour. The southern lowland comprises the main farming region. Soil and climatic conditions are favourable, and it is close to the country's largest market, Reykjavík and environs.



MAP INDEX

Political subdivisions

Austurland ..... 65 00 N 15 00 W
Höfudhborgarsvædhi ..... 64 15 N 21 30 W
Nordhurland Eystra ..... 65 30 N 17 00 W
Nordhurland Vestra ..... 65 15 N 19 45 W
Sudhurland ..... 64 15 N 19 30 W
Sudhurnes ..... 63 55 N 22 20 W
Vestfirdhir ..... 65 45 N 22 20 W
Vesturland ..... 64 45 N 21 30 W

Cities and towns

Akranes ..... 64 19 N 22 06 W
Akureyri ..... 65 40 N 18 06 W
Blönduós ..... 65 40 N 20 18 W
Bolungavik ..... 66 09 N 23 15 W
Borgarnes ..... 64 32 N 21 55 W
Dalvik ..... 65 58 N 18 32 W
Djúpivogur ..... 64 39 N 14 17 W
Egilsstadir ..... 65 16 N 14 25 W
Eskifjördhur ..... 65 04 N 14 00 W
Eyrarbakki ..... 63 52 N 21 09 W
Fáskrúdsfjördhur . 64 55 N 14 03 W
Grindavik ..... 63 50 N 22 26 W
Grundarfjördhur .. 64 55 N 23 15 W
Hafnarfjördhur ... 64 04 N 21 57 W
Hella ..... 63 50 N 20 24 W
Hofdhakaupstadhur, see Skagaströnd
Höfn ..... 64 15 N 15 13 W
Hólmavik ..... 65 43 N 21 41 W
Húsavik ..... 66 03 N 17 21 W
Hvammstangi ..... 65 24 N 20 57 W
Hveragerdhi ..... 64 00 N 21 12 W
Hvolsvöllur ..... 63 45 N 20 14 W
Innri-Njardhvik, see Njardhvik
Isafjördhur ..... 66 05 N 23 09 W
Keflavik ..... 64 01 N 22 34 W
Kópavogur ..... 64 06 N 21 55 W
Neskaupstadhur .. 65 09 N 13 42 W
Njardhvik (Innri-Njardhvik) . 63 58 N 22 31 W
Ólafsfjördhur .... 66 04 N 18 39 W
Ólafsvik ..... 64 53 N 23 43 W
Patreksfjördhur, see Vatneyri
Raufarhöfn ..... 66 27 N 15 57 W
Reykjavik ..... 64 09 N 21 57 W
Sandgerdhi ..... 64 03 N 22 42 W
Saudhárkrókur .... 65 45 N 19 39 W
Selfoss ..... 63 56 N 21 00 W
Seltjarnarnes .... 64 08 N 22 00 W
Seydhisfjördhur .. 65 16 N 14 00 W
Siglufjördhur .... 66 09 N 18 55 W
Skagaströnd (Hofdhakaupstadhur) ..... 65 50 N 20 19 W
Stykkishólmur ... 65 04 N 22 44 W
Thingeyri ..... 65 52 N 23 30 W
Thorlákshöfn .... 63 51 N 21 22 W
Thórshöfn ..... 66 12 N 15 20 W
Vatneyri (Patreksfjördhur) ..... 65 35 N 23 59 W
Vestmannaeyjar .. 63 26 N 20 16 W
Vík ..... 63 25 N 19 01 W
Vopnafjördhur .... 65 45 N 14 50 W

Physical features and points of interest

Askja Volcano .... 65 03 N 16 48 W
Atlantic Ocean .. 63 15 N 19 00 W
Axar Bay ..... 66 15 N 16 45 W
Bárdhar, Mount .. 64 38 N 17 32 W
Blanda, river .... 65 39 N 20 18 W
Breidha Bay .... 65 15 N 23 15 W
Denmark Strait .. 66 30 N 24 00 W
Detti Falls ..... 65 49 N 16 24 W
Dranga Glacier .. 66 09 N 22 15 W
Eyja Fjord ..... 65 54 N 18 15 W
Faxa Bay ..... 64 24 N 23 00 W
Fontur Point .... 66 23 N 14 32 W
Geysir Spring .... 64 19 N 20 18 W
Gils Fjord ..... 65 26 N 21 50 W
Glettinga Point .. 65 30 N 13 37 W
Greenland Sea .. 66 30 N 18 00 W
Grímsey, island .. 66 33 N 18 00 W
Gull Falls ..... 64 20 N 20 08 W
Heimaey, island . 63 26 N 20 17 W
Hekla Volcano .... 64 00 N 19 40 W
Hofs Glacier .... 64 49 N 18 48 W
Hofsá, river ..... 65 43 N 14 48 W
Hóp Lagoon .... 65 32 N 20 30 W
Húna Bay ..... 65 50 N 20 50 W
Hvannadals, Mount ..... 64 01 N 16 41 W
Hvítá, river .... 64 00 N 20 58 W
Hvítár, Lake .... 64 37 N 19 50 W
Ingólfs Headland . 63 48 N 16 39 W
Isa Fjord ..... 66 10 N 23 00 W
Jökulsá á Brú, river ..... 65 40 N 14 16 W
Jökulsá á Fjöllum, river .... 66 02 N 16 27 W
Jökulsá Canyon National Park .... 66 00 N 16 30 W
Lagar, river .... 65 40 N 14 18 W
Lang Glacier .... 64 39 N 20 12 W
Markar, river .... 63 32 N 20 05 W
Mýrdals Glacier .. 63 40 N 19 06 W
Mývatn, Lake .... 65 36 N 17 00 W
Ódádha Lava Field ..... 65 09 N 17 00 W
Öræfa Glacier .... 64 02 N 16 39 W



Arnarstapi, fishing village on Faxa Bay, Snæfell Peninsula, on the west coast of Iceland.
© Brian A. Vikander/West Light

Drainage and soils. Heavy rainfall feeds the numerous rivers and lakes in the glaciated landscape. Many of the lakes are dammed by lava flows or glacial ice. The presence of waterfalls is typical of the geologically young mountain landscape. The rivers are mainly debris-laden streams of glacial origin or clear streams formed by rainfall and springs of underground water. In the regions not drained by glacier rivers, fjords and smaller inlets cut into the rocky coasts. Because glacial erosion has often deepened the inner portions of the fjords, there are many fine natural harbours. Elsewhere the coasts are regular, sandy, and lined extensively with offshore sandbars that form lagoons to the landward side.

Iceland has soils of both mineral and organic composition. The mineral soils are basically a yellow-brown loess, formed by deposits of wind-transported matter. Both types of soil are suitable for agriculture, but, because of the slow rate of biological activity in the northern climate, they require heavy fertilization.

Climate. The climate of Iceland is maritime subarctic. It is influenced by the location of the country on the broad boundary between two contrasting air currents, one of polar and the other of tropical origin. The climate is affected also by the confluence of two ocean currents, the Gulf Stream, from near the Equator, and the East Greenland Current. The latter sometimes carries Arctic drift ice to Iceland's northern and eastern shores.

Effects of ocean currents

Seasonal shifts in temperature and precipitation are largely the result of weather fronts crossing the North Atlantic. Relatively cold weather, particularly in the northern part of the country, results from the movement of a front south of Iceland; mild, rainy weather is brought by the movement of a front northeastward between Iceland and Greenland. Although its northernmost points nearly touch the Arctic Circle, Iceland is much warmer than might be expected.

Mean annual temperatures for Reykjavík (on the west coast), Akureyri (in the north), and Kirkjubaejarklaustur (in the south) are, respectively, 40°, 38°, and 40° F (4°, 3°, and 4° C). For the same locations, the mean January temperatures are 31°, 28°, and 32° F (−0.5°, −2°, and 0° C), and the mean July temperature is 51° F (11° C). Snow falls about 100 days per year in the northwest, about 40 in the southeast. Annual precipitation ranges from 16 inches (410 millimetres) on some high northern plateaus to more than 160 inches on the southern slopes of some ice-capped mountains. In the south it averages about 80 inches. Gales are frequent, especially in winter, and occasionally heavy fogs may occur, but thunderstorms are rare. Reykjavík averages nearly 1,300 hours of bright sunshine a year. Often the aurora borealis appears, especially in fall and early winter.

Plant and animal life. Iceland lies on the border between a tundra vegetation zone of treeless plains and a taiga zone of coniferous forests. Only about one-fourth of the country is covered by a continuous carpet of vegetation. Bogs and moors are extensive, and sparse grasslands are often overgrazed. The remains of large birch forests are found in many places. A reforestation program instituted by the government in the 1950s has shown considerable success since the mid-1970s.

Foxes were the only land mammals in Iceland at the time of its settlement. Humans brought domestic and farm animals and accidentally introduced rats and mice. Later reindeer were introduced, and many are still found in the northeastern highlands. After 1930 mink that were brought in for the production of furs also became wild in the country. Birdlife in Iceland is varied. Many nesting cliffs are densely inhabited, and the colony of ducks at Lake Mývatn, in the north, is the largest and most varied in Europe. Salmon and trout abound in the lakes, brooks, and rivers. The fishing banks off the Icelandic shores are abundantly endowed with fish, although these resources have been considerably eroded by overexploitation. There are no reptiles or amphibians in Iceland.

Introduction of animal life

Settlement patterns. Because agriculture was the chief economic activity, the population of Iceland was evenly distributed throughout the inhabitable parts of the country until the end of the 19th century. With the advent of the fishing industry, commerce, and services at the beginning of the 20th century, the population became increasingly concentrated in towns and villages. By the end of the century, more than 90 percent of the population lived in communities of 200 or more people.

The mainstay of most coastal towns is fishing and fish processing. The greatest population concentration is in Reykjavík and its environs, with about three-fifths of Iceland's total population. Reykjavík is a modern, cosmopolitan urban centre that—in addition to being the seat of government—is the national focus of commerce, industry, higher education, and cultural activity. Akureyri, a fishing and educational centre of about 15,000 inhabitants situated on the Eyja Fjord in the north, is second in importance. Reykjanesbaer is a fishing port on the southwestern peninsula near Keflavík International Airport. The Vestmanna (Westman) Islands, off the southern coast, have some of the most important fishing operations in Iceland. Akranes, located across the bay from Reykjavík, is a service town for its region and has some industry. Ísafjördhur is a service town for the western fjord area. Seydhisfjördhur and Neskaupstadhur, on the eastern coast, are important ports for herring and capelin fishing. Höfn, on the southeastern coast, is also an important fishing port. Selfoss is in the southern lowlands, serving the farming region, and is the largest inland rural community in Iceland.

THE PEOPLE

The population of Iceland is extremely homogeneous. The inhabitants are descendants of settlers that began arriving

Norse and Celtic background

in AD 874 and continued in heavy influx for about 60 years thereafter. Historians differ on the exact origin and ethnic composition of the settlers but agree that between 60 and 80 percent of them were from Norway. The rest, from Scotland and Ireland, were largely Celtic. The dominant language in the period of settlement was Old Norse, the language spoken in Norway at the time. Through the centuries it has evolved into modern Icelandic, which is used throughout the country. There are no ethnic distinctions. The early Nordic and Celtic groups have long since merged, and the small number of subsequent immigrants have had no major effect on the population structure. The Lutheran faith has been the dominant religion since the mid-16th century. About nine-tenths of the population belongs to the state-supported Evangelical Lutheran church. There is freedom of religion.

The first comprehensive census in Iceland was taken in 1703, at which time 50,358 people were reported. The 18th century was marked by great economic hardship, and by 1801 the population had declined to 47,240. There began a slow increase in the 19th century, and by 1901 the population had risen to 78,470. Accelerated economic growth during the early decades of the 20th century was paralleled by a rapid growth in population, which in 1950 reached 143,973. During World War II and the early postwar period there was rapid improvement in the standard of living and a new acceleration in the rate of population growth. The annual growth rate reached its peak during the 1950s; it has been declining since 1960, primarily because of a sharply reduced birth rate and continued emigration. For a brief period from the late 1980s to the mid-1990s the birth rate rose again before resuming its downward trend. In the late 1980s the population reached a quarter of a million.

Between 1870 and 1914 there was large-scale emigration to Canada and the United States because of unfavourable conditions in Iceland; during that period emigrants outnumbered immigrants by the equivalent of about one-fifth of the 1901 population. Since 1901 emigration has continued to exceed immigration, though usually by only a small margin.

THE ECONOMY

The Icelandic economy is based heavily on fishing and the production of a broad variety of fish products, but it also includes manufacturing and services. Exports account for about two-fifths of the gross national product. Despite Iceland's small population, the economy is modern, and the standard of living is on a par with that of other European countries.

Most of Iceland's production is in private hands. Government ownership has declined since the early 1990s through increased privatization of government-owned enterprises. The state still owns two commercial banks and several other financial institutions and shares ownership of most electricity-generating systems with local governments. The central government receives a major portion of its income from a value-added tax and a progressive income tax, whereas local governments derive most of their revenue from a flat-rate income tax and property levies.

Economic growth rates

Since World War II the government has aimed for a high rate of economic growth and full employment, and fluctuations in fish prices and catches have been an important influence on the economy. Iceland's real gross domestic product (GDP) increased by an average of about 4 percent per year after the war. After 1987, however, there was a slowdown in economic growth owing to limits imposed on fish catches in response to the depletion of fish stocks that had been overexploited for many years. From the late 1980s to the late 1990s the annual GDP growth rate averaged less than half what it had been. From 1994, however, there was a strong resumption of growth, mainly as a result of an improving fish catch. The inflation rate was high up to the end of the 1980s but thereafter declined. A low rate of inflation did not become a priority until the early 1990s, but Iceland now enjoys as low a rate as the other northern European countries. Unemployment has remained low.

Resources. Iceland's energy resources are vast. Feasible hydroenergy is estimated at nearly six gigawatts and geothermal energy at more than 1.5 million gigawatt hours per year. Only about one-eighth of the hydroelectric power of the country's rivers has been tapped. Geothermal energy heats all of Reykjavík and several other communities; it provides steam for industrial energy and is used in commercial vegetable farming in greenhouses.

Fishing. A steady improvement in Iceland's fishing technology has increased catches despite the gradual erosion of what once were enormously rich fish populations off the country's coasts. During the late 1980s and early 1990s the concern over declining fish stocks led the government to strengthen already strict catch quotas to further husband eroded fish stocks—particularly cod, the most important species. The strict quota regime paid off with a sharp increase in the cod stock in the late 1990s. Those catch quotas for domestic waters led to increased fishing in foreign waters, particularly in the Barents Sea and off the coast of Newfoundland.

Cod and capelin make up about two-thirds of the total catch, and such whitefish (demersal) species as cod and haddock are exported fresh, frozen, salted, or dried. The capelin and herring catches usually are reduced to oil and meal but also are salted. In the mid-1990s Iceland's total fish catch was between about 1.5 million and 2 million tons, of which about one-third was whitefish species and two-thirds were capelin and herring. Such fishing-related industries as boatyards, repair docks, and net factories are also important.

Agriculture. As is the case throughout the Nordic countries, less than 5 percent of Iceland's population is engaged in agriculture, and this number continues to decline. The raising of livestock—mostly sheep—and dairy farming are the main occupations. About one-fifth of the land is arable, most of it used for grazing. Greenhouses are common, especially in the southern part of the country. Iceland is virtually self-sufficient in fresh foods and dairy items, but it imports most other foodstuffs.

Dairy farming

Industry. The main manufacturing enterprise for export is aluminum production, which uses domestic hydroelectricity to smelt aluminum from imported alumina. Other manufactured goods for export include ferrosilicon, an alloying agent for steel production; diatomite, an industrial filtration agent produced from diatomaceous earth with geothermal steam; fish-processing equipment; fishing gear; and prosthetic devices. There are also small industries that produce computer software, cement, fertilizer, food, clothing, and books.

Finance. Iceland has three commercial banks with branches throughout the country and a number of savings banks. Other financial institutions include investment credit funds, private pension funds, insurance companies, and securities firms. The financial sector was gradually deregulated in the 1980s. Interest rates were left to market forces, an important stock and bond market developed, and monetary policy—previously quite inflationary—became comparable to that of other industrialized countries. The reform of the financial market played an important part in bringing inflation under control. Capital movements to and from other countries were almost completely liberalized by the mid-1990s. The exchange rate of the country's currency, the króna, is linked to the U.S. dollar, the Japanese yen, and the euro, the currency of the European Union.

Transportation and tourism. The historic isolation of Iceland, caused by the rough seas of the North Atlantic and the country's small market and industry, was broken when steam vessels began to visit Icelandic shores late in the 19th century. The first telegraph cable to Iceland was laid in 1906, and the Iceland Steamship Company (Eimskip) was founded in 1914. Before the 20th century roads were practically unknown, the horse being the means of transportation throughout the island. Iceland has no railroads. Domestic transportation is provided by truck, car, plane, or coastal boats. Most of Iceland's main rural roads are paved, as are most streets in towns and villages. The majority of minor country roads, however, are still gravel. During the summer driving is possible on the extensive sandy plains in the uninhabited interior, permitting expeditions between the glaciers. The Hringvegur ("Ring Road")

stretches for about 875 miles (1,400 kilometres), forming a circle around the island. The merchant marine fleet transports most of Iceland's imports and exports. Icelandair (Flugleidir), a major international carrier, as well as local air service carriers are important internally in compensating for the limited road system. Keflavík International Airport, the country's primary gateway, is located about 30 miles (48 kilometres) west of Reykjavík. Icelandair has helped make the tourist trade increasingly important to the national economy. Foreign tourists number more than 200,000 a year, and the tourist industry is an important earner of foreign exchange.

Icelandair and tourism

Foreign trade. More than three-fifths of Iceland's exports go to the European Union (EU), which also is responsible for more than half of Iceland's imports. About one-eighth of exports go to the United States and about one-tenth to Japan. Iceland has been a member of the European Free Trade Association (EFTA) since 1970. It has stopped short of applying for membership to the EU because of its concern that the EU would control its fishing resources.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Iceland is a parliamentary democracy with a directly elected president as head of state. The powers of the president are similar to those of other heads of state in western European democracies. Real power rests with the 63-member parliament, the Althing (Althingi). One of the oldest legislative assemblies in the world, the Althing is a unicameral legislature in which members serve four-year terms unless parliament is dissolved. The executive branch is headed by a cabinet that must maintain majority support in parliament—or at least avoid censure—otherwise it must resign. The judiciary consists of a supreme court and a system of lower courts, most of which hear both civil and criminal cases. Cases are heard and decided by appointed judges; there is no jury system. Citizens are guaranteed the civil rights customary in Western democracies.

Local government in Iceland is chiefly responsible for primary education, municipal services, and the administration of social programs. The country is divided into 124 municipalities. Since the 1970s their number has decreased by nearly half as a result of consolidation. Each municipality administers local matters through an elected council.

The president, Althing, and local councils are elected every four years, but not necessarily all at once. All citizens 18 years of age and older may vote. Since the late 1970s the Independence Party, centre to conservative in political outlook, has commanded about one-third of the popular vote. The Progressive Party, which has captured the second largest proportion of the vote during this period, draws its strength from rural areas. Among the other main political parties are the Social Democratic Party, the People's Alliance, and the Women's Alliance.

Foreign relations. Iceland joined the United Nations in 1946, a year after its founding, and is a charter member of the North Atlantic Treaty Organization (NATO). In the post-World War II period it has based its foreign policy on peaceful international cooperation and participated in joint Western defense efforts. It does not maintain armed forces. The United States, having assumed responsibility for Iceland's defense, maintains a naval air station at Keflavík International Airport under NATO auspices. Iceland's entry into distant fishing waters has caused friction with other nations over fishing rights. Fishing in the Barents Sea is a source of contention with Norway and Russia, as is herring fishing in the open ocean between Norway and Iceland. Moreover, Canada has objected to Iceland's shrimp fishing off the coast of Newfoundland.

Health and welfare. Iceland, with compulsory health insurance that finances most medical services, has a high standard of public health and one of the highest life expectancies in the world. Hospital inpatient services are provided entirely without charge, other medical services at low cost. Dental care is partially subsidized for children up to age 16 and for retirees with low incomes. Heart disease and cancer together account for about one-half of all deaths. Welfare services include unemployment insurance, old-age and disability pensions, family and childbearing allowances, and sickness benefits. The medical and welfare systems are financed through taxation by central and local government.

Comprehensive health insurance

Education. Almost all schools from the primary level through the university are free. Education is compulsory through age 16, and secondary and higher education is widely available. Students can enroll in four-year academic colleges at the age of 15 or 16. Graduation from one of these colleges entitles the student to admission to the University of Iceland, founded in 1911, in Reykjavík. A second university was established at Akureyri in 1987. There are also a number of technical, vocational, and specialized schools.

CULTURAL LIFE

Icelanders are proof that a rich cultural life can be developed despite a small population. The country's literary heritage stems from writers of the 12th to 14th centuries who vividly recorded the sagas of Iceland's first 250 years. Other traditional arts include weaving, silver crafting, and wood carving. Poetry was the great literary form of expression in the 19th century, whereas the novel and drama were the prime forms of literature in the 20th century.

The Reykjavík area, which supports several professional theatres, a symphony orchestra, an opera, and a number of art galleries, bookstores, cinemas, and museums, has a cultural environment that compares favourably with those of cities several times its size. It also holds biennial international art festivals.

The arts. Art in Iceland was long connected with religion, first the Roman Catholic church and later the Lutheran church. The first professional secular painters appeared in Iceland in the 19th century. Gradually increasing in number, these painters, such as Jóhannes Kjarval, highlighted the character and beauty of their country. Painting continues to thrive in Iceland, where artists have fused foreign influences with local heritage. The old traditions in silver working have been retained, the most characteristic of which is the use of silver thread for ornamentation.

Literature is also alive and well in Iceland. The literary tradition of the saga has been revived, and Iceland was often the setting of 20th-century fiction. Several Icelandic writers have received international acclaim, such as Halldór Laxness, who received the Nobel Prize for Literature in 1955. Other native writers have written for the theatre, and their work has grown more international in theme and setting. Music also enjoyed a tremendous upsurge after World War II. The programs of the Iceland Symphony are drawn from a classical repertoire and the work of modern Icelandic composers, and one or more operas or musicals are performed every year at the National Theatre and the Icelandic Opera. Popular music by Icelandic performers, such as Björk, has gained international commercial success and critical acclaim.

Literary activity

National folk traditions in applied art have achieved a new popularity. Old designs and forms have been revived, some modified to please modern tastes. Wool, knitted or woven, is the most commonly used material. Many people in the country participate in this industry, creating high-quality goods.

Sports. Glíma, a form of wrestling that originated with the Vikings, is still practiced in Iceland. Swimming in naturally heated pools, pony trekking, and various ball games also are popular. Team handball became the national sport in the 1980s, with Iceland's national team ranked among the top teams in the world.

Cultural institutions. The National Library of Iceland (founded in 1818) and the University Library (1940) merged in 1994. The National Archives were founded in 1882. The National Museum of Iceland, dating from 1863, has collections representing native Icelandic culture beginning in the Viking Age. Many old houses and ruins throughout the country are preserved under its auspices. The National Gallery of Iceland was founded in 1884, and the great majority of its works are by modern Icelandic artists. The Natural History Museum was founded in 1889.

The National Theatre began operation in 1950. It performs Icelandic as well as foreign classical and modern plays, operas, ballets, and musicals. The Reykjavík Theatre is the other full-time professional repertory theatre. Sev-

eral theatre groups present numerous plays and musicals, both in Reykjavík and the countryside.

Book publishing also is an active Icelandic tradition. More than 1,000 book titles are published every year. There are three daily newspapers in the country and numerous magazines and journals. The Icelandic Literary Society, founded in 1816, specializes in the publication of historical and classical works. (Va.K./B.M.)

For statistical data on the land and people of Iceland, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

EARLY HISTORY

Settlement (c. 870–c. 930). Iceland apparently has no prehistory. According to stories written down some 250 years after the event, the country was discovered and settled by Norse people in the Viking Age. The oldest source, *Íslendingabók* (*The Book of the Icelanders*), written in about 1130, sets the period of settlement at about AD 870–930. The other main source, *Landnámabók* (*The Book of Settlements*), of 12th-century origin but known only in later versions, states explicitly that the first permanent settlers, Ingólfr Arnarson and his wife, Hallveig Fróðadóttir, came to Iceland to settle in the year 874. They came from Norway and chose as their homestead a site that Ingólfr named Reykjavík. *The Book of Settlements* then enumerates more than 400 settlers who sailed with their families, servants, and slaves to Iceland to stake claims to land there. Most of the settlers came from Norway, but some came from other Nordic countries and from the Norse Viking Age settlements in the British Isles.

Sources

Some scholars have found it difficult to believe that such a large and habitable country remained unpopulated until the 9th century. On the whole, however, archaeological finds support the documentary evidence and place Iceland among Norse Viking Age settlements of the late 9th or early 10th century. The Icelandic language testifies to the same origin: Icelandic is a Nordic language and is most closely related to the dialects of western Norway.

Although the island was not populated until the Viking Age, Iceland probably had been known to people long before that time. The 4th-century-BC Greek explorer Pytheas of Massalia (Marseille) described a northern country that he called Thule, located six days' sailing distance north of Britain. In the 8th century, Irish hermits who had begun to sail to Iceland in search of solitude also called the island Thule. It is unknown, however, if Pytheas and the hermits were describing the same island. According to the early Icelandic sources, some Irish monks were living in Iceland when the Nordic settlers arrived, but these monks soon left because they were unwilling to share the country with heathens.

Commonwealth (c. 930–1264). At the time of Iceland's settlement, Norse people worshiped gods whom they called *æsir* (singular *ás*), and this religion has left behind an extensive mythology in Icelandic literature. Thor seems to have been the most popular of the pagan gods in Iceland, although Odin is thought to have been the highest in rank. It appears that heathen worship was organized around a distinct class of priest-chieftains, called *godar* (singular *godi*), of which there were about 40. In the absence of royal power in Iceland, the *godar* were to form the ruling class in the country.

The *godar*

By the end of the settlement period, a general Icelandic assembly, called Althing, had been established and was held at midsummer on a site that came to be called Thingvellir. This assembly consisted of a law council (*lögrétta*), in which the *godar* made and amended the laws, and a system of courts of justice, in which householders, nominated by the *godar,* acted on the panels of judges. At the local level, three *godar* usually held a joint assembly in late spring, at which a local court operated, again with judges nominated by the *godar.* All farmers were legally obliged to belong to a chieftaincy (*godord*) but theoretically were free to change their allegiance from one *godi* to another; the *godar* were allotted a corresponding right to expel a follower. Some scholars have seen in this arrangement a resemblance to the franchise in modern societies. On the other hand, there was no central authority to ensure that the farmers would be able to exercise their right in a democratic way. No one was vested with executive power over the country as a whole. In any case, no trace of democratic practice reached farther down the social scale than to the heads of farming households: women and workers (either free or enslaved) had no role in the political system.

Christianization. By the end of the 10th century the Norwegians were forced by their king, Olaf I Tryggvason, to accept Christianity. The king also sent missionaries to Iceland, who according to 12th-century sources were highly successful in converting the Icelanders. In 999 or 1000 the Althing made a peaceful decision that all Icelanders should become Christians. In spite of this decision, the *godar* retained their political role, and many of them probably built their own churches. Some were ordained, and as a group they seem to have closely controlled the organization of the new religion. Two bishoprics were established, one at Skálholt in 1056 and the other at Hólar in 1106. Literate Christian culture also transformed lay life. Codification of the law was begun in 1117–18. A little later the Icelanders began to write their sagas, which were to reach their pinnacle of literary achievement in the next century.

Economic life. Historians believe that early Icelandic society was a prosperous one. The country proved to be well suited for sheep and cattle, and both were raised for meat and milk. The sheep also yielded wool, and homespun cloth became the chief export. There was some agriculture, but grain was always imported. Timber was also imported: the only indigenous wood was dwarf birch. However abundant driftwood may have been, it could not satisfy the needs of the whole population. The Icelanders built large turf-clad houses on bulky timber frames, and some of the churches were built entirely of timber.

In spite of the seeming abundance, the end was coming for an independent Icelandic commonwealth. In Norway, royal power gained strength in the early 13th century, and the king set himself the aim of uniting all Norwegian Viking Age settlements under his reign. By that time, a score of powerful *godar,* belonging to some five families, held almost all the chieftaincies in Iceland. The mid-13th century is characterized by a bloody struggle for power among these chieftains. Finally, in 1262–64 all Icelandic chieftains and representatives of the farmers were persuaded to swear allegiance to the king of Norway, partly in the hope that he would bring peace to the country.

End of the commonwealth

ICELAND UNDER FOREIGN RULE

Late Middle Ages (1264–c. 1550). To a large extent, Iceland was ruled separately from Norway. It had its own law code, and the Althing continued to be held at Thingvellir, though mainly as a court of justice. Most of the royal officials who succeeded the chieftains were Icelanders. In 1380 the Norwegian monarchy entered into a union with the Danish crown, but that change did not affect Iceland's status within the realm as a personal *skattland* ("tax land") of the crown.

Economic growth and decline. A fundamental change in Iceland's economy took place in the early 14th century when Norwegian merchants began to import dried fish from Iceland to Bergen, Nor. English merchants in Bergen became acquainted with Icelandic fish supplies, and shortly after 1400 they themselves began sailing to Iceland and either catching fish or buying it from local fishermen. The Danish crown repeatedly tried to stop English trade in Iceland but lacked the naval power with which to defend its remote possession. One of the royal governors was killed by the English when he tried to stop their trade, an event that led indirectly to clashes between Denmark and England (1468–73). In the early 16th century, English interest in Iceland declined, partly because rich fishing grounds had been discovered off the North American coast of Newfoundland. Instead, Germans became the chief foreigners to fish and trade in Iceland.

Importance of fishing

In spite of the rise of a profitable export industry, it is generally believed that Iceland's economy deteriorated in the late Middle Ages. The birchwood that had covered great parts of the country was gradually depleted, in part

because it was excellent for making charcoal. The destruction of the woodland, together with heavy grazing, led to extensive soil erosion. The climate also became more severe, and grain growing was given up altogether. At the same time, more and more of the land was acquired by ecclesiastical institutions and wealthy individuals, to whom the farmers had to pay rent.

The Reformation. The Lutheran Reformation, which was instituted in Denmark in the 1530s, met greater resistance in Iceland than anywhere else in the realm. In 1541 the bishop of Skálholt was captured by the governor, and Lutheranism was introduced in his diocese, which covered three-quarters of the country. In the northern diocese of Hólar, Bishop Jón Arason was to hold out against Lutheranism for a decade longer. In 1550 he was finally captured and beheaded, without benefit of law or clergy, and all resistance to the Reformation ended. Jón Arason's death is traditionally taken to mark the end of the Middle Ages in Iceland.

Growth of Danish royal power (c. 1550–c. 1830). After the Reformation the royal treasury confiscated all lands that had belonged to the Icelandic monasteries. Also during the 16th century the German traders were ousted. In 1602 all foreign trade in Iceland was monopolized by a royal decree and handed over to Danish merchants, who paid a rent on it to the crown. This arrangement lasted intact until 1787, when the monopoly was abolished. Only subjects of the Danish crown, however, were permitted to carry on foreign trade, a restriction that remained in force until 1855.

Danish monopoly of trade

On the constitutional level, also, the Danish crown increased its hold on Iceland, at least in formal terms. In 1661 Frederick III introduced an absolute monarchy in Denmark and Norway, and in the following year his absolutism was acknowledged in Iceland. This event was not of any great immediate significance in Iceland; local officials, most of whom were Icelanders, continued to make important political decisions. Danish officials in Copenhagen rarely had enough knowledge of or interest in Icelandic affairs to enforce their will if the Icelandic officials in Iceland were unanimous on a different policy.

Nevertheless, the bureaucratic state, which formed the backbone of absolutism, was gradually introduced into Iceland. An essential part of that development was the emergence of a town nucleus in Reykjavík, the first one in this hitherto entirely rural country. In the 1750s a tiny village grew up in Reykjavík as a result of a semiofficial attempt to start a wool-processing factory there. Within half a century the two ancient bishoprics were united, with the bishop residing in Reykjavík. The Althing was abolished in 1800, and an appeals court was set up in Reykjavík to succeed it. A few years later the Danish governor also settled in the town, which by then had about 300 inhabitants.

Economic conditions

Only since the 18th century has there been sufficient information available to allow for a reliable survey of Iceland's economy. The first census, for example, was held in 1703, revealing that the population was just over 50,000. The main occupation of most of the inhabitants was farming, but an important auxiliary occupation, undertaken mostly by rural labourers on the southern and western coasts in late winter and spring, was fishing. With few exceptions, labourers were obliged to stay in the domestic service of a farmer, and the establishment of permanent households in fishing stations was severely restricted. Thus, the landowners, with most of the native officials in their number, succeeded in monopolizing fishing, and they prevented it from becoming an independent industry.

The 18th century, a time of prosperity elsewhere in the Danish realm, was a period of decline and increasing poverty in Iceland. In the 1780s the country was plagued by famine—caused by a volcanic eruption and subsequent years of cold weather—which killed one-fifth of the population. In spite of these hardships, it appears that few Icelanders were critical of the status of the country within the Danish realm. In 1809 a Danish adventurer, Jørgen Jørgensen, seized power in Iceland for two months. When he was removed and Danish power was restored, he received no support from the Icelandic population. Five years later, when Norway was united politically with Sweden, no Icelander expressed a wish to follow suit.

MODERN ICELAND

Struggle for independence (c. 1830–1904). In the 1830s the Danish crown eased its absolutism by setting up four consultative diets in the realm. Iceland was allotted two representatives in one of them, the assembly of the Danish Isles meeting at Roskilde, Den. This arrangement was never popular in Iceland, and the adherents to emerging nationalist and liberal sentiments wanted the Icelandic Althing to be restored on its ancient site as a consultative assembly for the nation. In 1840 Christian VIII decided to grant them their wish, and five years later a restored Althing, vested with consultative power, met for the first time, not at Thingvellir, however, as originally intended, but in Reykjavík. Franchise to the assembly was almost entirely restricted to officials and (male) farmers.

In 1848, only three years after the first convention of the Althing, Frederick VII renounced his absolute power, and a constitutional assembly was summoned to prepare a representative democracy in Denmark. This led inevitably to the question of what was to become of Iceland in the new form of government. By this time, Iceland had a relatively undisputed political leader, Jón Sigurdhsson, a philologist living in Copenhagen. Sigurdhsson argued that the king could only give his absolute rule over Iceland back to the Icelanders themselves, since they were the ones who had surrendered it to him in 1662. This claim was met with a royal pledge that the constitutional status of Iceland would not be decided until the Icelanders had discussed the matter at a special assembly in Iceland. This assembly met in 1851, but no agreement could be reached between the Icelandic representatives and the Danish government. The assembly was dissolved in disappointment. A stalemate of more than 20 years ensued, but the Althing decided to use the occasion of the millennium of Iceland's settlement (1874) to accept the status that Danish authorities were by then willing to grant. Thus, in 1874 the king presented Iceland with a constitution whereby the Althing was vested with legislative power in internal affairs. As before, however, the cabinet minister responsible for Iceland was the minister of justice in the Danish government.

Constitution of 1874

For an additional three decades the Icelanders fought to get executive power transferred to Iceland. In 1901 the path was opened when rule by parliamentary majority was introduced in Denmark and the Liberal Party—always more positive than the Conservatives toward the Icelanders—came into power. In 1904 Iceland got home rule, and the first Icelandic minister opened his office in Reykjavík. At the same time, rule by parliamentary majority was introduced.

The high level of political activity in 19th-century Iceland stands in sharp contrast with its economic stagnation. The considerable growth of Iceland's population put increasing strain on the badly eroded rural areas, and for many people the only visible solution was emigration to North America. Some 15,000 Icelanders emigrated between 1870 and 1914, most of them to Canada. Virtually the only successful technical innovation during that period was the introduction of decked fishing vessels, which made it possible to catch fish farther offshore than could be done on open boats. Still, at the beginning of the 20th century, more than half of the annual catch was still taken in open boats.

Home rule and sovereignty (1904–44). The period of home rule (1904–18) was one of rapid progress. Motors were installed in many of the open fishing boats, and a number of steam-driven trawlers were acquired. The country was connected by telephone cable with Europe. School attendance was made compulsory for children in towns and villages, and a number of schools were built. The University of Iceland was established in Reykjavík, which by 1918 had a population of 15,000. All restrictions on the freedom to move to the fishing villages were either abolished or quietly forgotten. There was a radical transformation in the occupational structure of the country, which, in turn, led to the advent of a labour movement. In 1916 a national organization of trade unions was es-

tablished. By then, unions were already widely accepted by employers as negotiating bodies, but their formal status was not legalized until 1938. In the political arena, democracy was extended to new groups in the society. Women and propertyless men were given the franchise, subject to certain qualifications, in 1915. Four years earlier, a law had been passed that gave women the right to attend schools of higher education, enter into the professions, and occupy any public office in the country.

The union

The struggle continued for greater autonomy until the dispute with Denmark was solved by an agreement. On Dec. 1, 1918, Iceland became a separate state under the Danish crown, with only foreign affairs remaining under Danish control. Either party, however, had the right to call for a review of the treaty, and if negotiations about its renewal proved fruitless at the end of 25 years (*i.e.*, 1943) it would be terminated.

The struggle for independence that had shaped Icelandic politics for almost a century now subsided, and in the 1920s a new system of political parties based on class divisions emerged. Class antagonism grew more severe during the Great Depression of the 1930s; the depression was prolonged in Iceland when the outbreak of the Spanish Civil War in 1936 closed the important Spanish market for Icelandic fish. The problem of high unemployment persisted until after the outbreak of World War II.

The German occupation of Denmark in April 1940 effectively dissolved the union between Iceland and Denmark. A month later, British forces occupied Iceland. In 1941 the United States took over the defense of Iceland and stationed a force of 60,000 in the country. The foreign forces brought employment, prosperity, and high inflation to the population, which then numbered about 120,000.

The war made it impossible for Iceland and Denmark to renegotiate their treaty. In spite of great resentment in Denmark, the Icelanders decided to terminate the treaty, break all constitutional ties with Denmark, and establish a republic. On June 17, 1944, now celebrated as National Day, the Icelandic republic was founded at Thingvellir, with Sveinn Björnsson as its first president.

The Icelandic republic. Since the prosperous years of World War II, Iceland has developed into a modern welfare state with growing production and consumption. A rapid restoration of the trawler fleet after the war prevented the return of prewar unemployment. Fish freezing became a highly technical industry and the mainstay of Iceland's exports. The economy became characterized by expansion, full employment, high inflation, and much unprofitable investment. It became normal to work overtime and for women to enter the labour market.

The tendency toward overexpansion—which seemed to have been checked in the 1990s—was caused in part by weak political leadership. Generally Iceland has been ruled by coalition governments, and these coalitions have tended to blend parties of the political left and right. This blurring of ideologies has probably been caused by another dividing line in Icelandic postwar politics: that between integrationism, generally espoused by the Independence and Social Democratic parties, and isolationism, identified with the Progressives and People's Alliance. This contrast has come to a head over three recurrent issues: defense, European integration, and extension of fishing limits.

From the time that Iceland joined the North Atlantic Treaty Organization (NATO) in 1949, the Independence Party has firmly supported NATO while the People's Alliance has been an ardent opponent. The Social Democratic Party and the Progressives also have supported NATO membership, though with some reluctance over the presence of American forces. Since the 1980s the issue of defense has yielded the foreground to Iceland's relations with Europe. Iceland entered the European Free Trade Association (EFTA) in 1970, in the period of an Independence and Social Democratic coalition, against the votes of the People's Alliance. Since then the Social Democratic Party alone has sought full Icelandic membership in the EU and thus appears to be the most integrationist party.

After World War II Iceland extended its exclusive fishing zone from 3 nautical miles (5.6 kilometres) in 1950 to 200 nautical miles (370 kilometres) in 1975. This action provoked strong protests, and the British navy was repeatedly sent to Icelandic fishing grounds to protect British trawlers. The confrontation, commonly known as the "Cod Wars," ended in 1976 when Britain recognized the 200-nautical-mile limit. Nevertheless, while all the political parties supported Iceland's claims, only the more isolationist parties were willing to risk Iceland's relations with its NATO partners. Eventually, as the fish stocks around Iceland began to be depleted, Icelandic fishing firms started deep-sea fishing on remote grounds. This led to disputes with other fishing nations—particularly with Norway and Russia over fishing in the Barents Sea.

The status of women has formed still another dimension of Icelandic politics. A woman, Vigdís Finnbogadóttir, served as president of the republic for four terms (1980–96), enjoying great popularity, and a Women's Alliance has been represented in the parliament since 1983. However, the Icelandic president typically is not influential in politics, and the Women's Alliance has yet to participate in forming a government. Furthermore, women still earn less income than men, suggesting that they have not yet obtained full equality. (Gu.K.)

For later developments in the history of Iceland, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 923 and 972.

BIBLIOGRAPHY

Physical and human geography. JÓHANNES NORDAL and VALDIMAR KRISTINSSON (eds.), *Iceland, the Republic* (1996), provides a comprehensive general survey. Geologic, geothermal, geographic, and ecological characteristics of the country are studied in ARI TRAUSTI GUDMUNDSSON and HALLDÓR KJARTANSSON, *Earth in Action: An Outline of the Geology of Iceland* (1996); BJÖRN H. RÓARSSON and SIGURDUR SVEINN JÓNSSON, *Geysers and Hot Springs in Iceland*, trans. from Icelandic (1992); THORLEIFUR EINARSSON, *Geology of Iceland: Rocks and Landscape* (1994; originally published in Icelandic, 1991); HÖRDUR KRISTINSSON, *A Guide to the Flowering Plants and Ferns of Iceland* (1987); and STURLA FRIDRIKSSON, *Surtsey* (1975). Illustrated descriptions and guidebooks include BJÖRN BÚRIKSSON, *The Beauty of Iceland* (1992); BERNARD SCUDDER and PÁLI STEFÁNSSON, *Iceland: Life and Nature on a North Atlantic Island* (1991); HJÁLMAR R. BÁRDARSON, *Ice and Fire: Contrasts of Icelandic Nature*, trans. from Icelandic, 4th ed. (1991); MAX SCHMID, *Iceland: The Exotic North* (1985); STEINDÓR STEINDÓRSSON FRÁ HLÖDUM (ed.), *Iceland Road Guide*, 4th updated ed. (1988); and DAVID WILLIAMS, *Essential Iceland* (1992).

Icelandic society, economy, culture, and political orientation are reviewed in *The Economy of Iceland* (annual), published by the Central Bank of Iceland; *OECD Economic Surveys: Iceland* (annual), published by the Organisation for Economic Co-operation and Development; GUDMUNDUR JÓNSSON and MAGNÚS S. MAGNÚSSON (eds.), *Icelandic Historical Statistics* (1997), a wide-ranging statistical abstract of Icelandic society; HJÁLMAR R. BÁRDARSON, *Iceland: A Portrait of Its Land and People*, 3rd ed., trans. from Icelandic (1982, reissued as 3rd ed., 1989); KRISTJÁN ELDJÁRN, *Ancient Icelandic Art* (1957); CAROL J. CLOVER and JOHN LINDOW (eds.), *Old Norse-Icelandic Literature: A Critical Guide* (1985); ÓLAFUR TH. HARDARSON, *Parties and Voters in Iceland: A Study of the 1983 and 1987 Elections* (1995); SIGURDUR A. MAGNÚSSON and VLADIMIR SICHOV, *Iceland Crucible: A Modern Artistic Renaissance* (1985); and GÍSLI PÁLSSON and E. PAUL DURRENBERGER, *Images of Contemporary Iceland: Everyday Lives and Global Contexts* (1996). (B.M.)

History. JÓN R. HJÁLMARSSON, *History of Iceland: From the Settlement to the Present Day* (1993), is a survey of the subject. GUDMUNDUR HÁLFDANARSON, *Historical Dictionary of Iceland* (1997), is useful and has a good bibliography of works in English. Scholarly studies include JÓN JÓHANNESSON, *A History of the Old Icelandic Commonwealth: Íslendinga Saga*, trans. from Icelandic (1974); and JESSE L. BYOCK, *Medieval Iceland: Society, Sagas, and Power* (1988, reissued 1993). E. PAUL DURRENBERGER and GÍSLI PÁLSSON (eds.), *The Anthropology of Iceland* (1989), includes essays covering the commonwealth period, 930–1262. Studies on the 20th century, with historical introductions, are GYLFI TH. GÍSLASON, *The Challenge of Being an Icelander*, trans. from Icelandic (1990); and ESBJÖRN ROSENBLAD and RAKEL SIGURDARDÓTTIR-ROSENBLAD, *Iceland from Past to Present* (1993; originally published in Swedish, 1990). GUDMUNDUR JÓNSSON and MAGNÚS S. MAGNÚSSON (eds.), *Icelandic Historical Statistics* (1997), is an extensive bilingual (Icelandic and English) survey. BJÖRN THORSTEINSSON, *Island*, trans. from Icelandic (1985), is a historical survey in Danish. SIGURDUR LÍNDAL (ed.), *Saga Islands* (1974–), is a comprehensive multivolume history in Icelandic. (Gu.K.)

Ideology

An ideology is a form of social or political philosophy in which practical elements are as prominent as theoretical ones; it is a system of ideas that aspires both to explain the world and to change it.

This article describes the nature, history, and significance of ideologies in terms of the philosophical, political, and international contexts in which they have arisen. For discussions of particular categories of ideology, see the *Macropædia* article SOCIO-ECONOMIC DOCTRINES AND REFORM MOVEMENTS, MODERN. Specific ideologies are described in the *Micropædia* under such headings as COMMUNISM and FASCISM.

The article is divided into the following sections:

Origins and characteristics of ideology 768
The philosophical context 768
  Ideology and religion
  Ideology in early political philosophy
  Hegel and Marx
  The sociology of knowledge
The political context 770
  Ideology, rationalism, and romanticism
  Ideology and terror
  Ideology and pragmatism
The context of international relations 771
  Ideology in the World Wars
  Ideology of the Cold War
Bibliography 772

ORIGINS AND CHARACTERISTICS OF IDEOLOGY

Science of ideas

The word first made its appearance in French as *idéologie* at the time of the French Revolution, when it was introduced by a philosopher, A.-L.-C. Destutt de Tracy, as a short name for what he called his "science of ideas," which he claimed to have adapted from the epistemology of the philosophers John Locke and Étienne Bonnot de Condillac, for whom all human knowledge was knowledge of ideas. The fact is, however, that he owed rather more to the English philosopher Francis Bacon, whom he revered no less than did the earlier French philosophers of the Enlightenment. It was Bacon who had proclaimed that the destiny of science was not only to enlarge man's knowledge but also to "improve the life of men on earth," and it was this same union of the programmatic with the intellectual that distinguished Destutt de Tracy's *idéologie* from those theories, systems, or philosophies that were essentially explanatory. The science of ideas was a science with a mission; it aimed at serving men, even saving them, by ridding their minds of prejudice and preparing them for the sovereignty of reason.

Destutt de Tracy and his fellow *idéologues* devised a system of national education that they believed would transform France into a rational and scientific society. Their teaching combined a fervent belief in individual liberty with an elaborate program of state planning, and for a short time under the Directory (1795–99) it became the official doctrine of the French Republic. Napoleon at first supported Destutt de Tracy and his friends, but he soon turned against them, and in December 1812 he even went so far as to attribute blame for France's military defeats to the influence of the *idéologues*, of whom he spoke with scorn.

Thus *ideology* has been from its inception a word with a marked emotive content, though Destutt de Tracy presumably had intended it to be a dry, technical term. Such was his own passionate attachment to the science of ideas, and such was the high moral worth and purpose he assigned to it, that the word *idéologie* was bound to possess for him a strongly laudatory character. And equally, when Napoleon linked the name of *idéologie* with what he had come to regard as the most detestable elements in Revolutionary thought, he invested the same word with all of his feelings of disapprobation and mistrust. Ideology was, from this time on, to play this double role of a term both laudatory and abusive not only in French but also in German, English, Italian, and all the other languages of the world into which it was either translated or transliterated.

Some historians of philosophy have called the 19th century the age of ideology, not because the word itself was then so widely used, but because so much of the thought of the time can be distinguished from that prevailing in the previous centuries by features that would now be called ideological. Even so, there is a limit to the extent to which one can speak today of an agreed use of the word. The subject of ideology is a controversial one, and it is arguable that at least some part of this controversy derives from disagreement as to the definition of the word *ideology*. One can, however, discern both a strict and a loose way of using it. In the loose sense of the word, ideology may mean any kind of action-oriented theory or any attempt to approach politics in the light of a system of ideas. Ideology in the stricter sense stays fairly close to Destutt de Tracy's original conception and may be identified by five characteristics: (1) it contains an explanatory theory of a more or less comprehensive kind about human experience and the external world; (2) it sets out a program, in generalized and abstract terms, of social and political organization; (3) it conceives the realization of this program as entailing a struggle; (4) it seeks not merely to persuade but to recruit loyal adherents, demanding what is sometimes called commitment; (5) it addresses a wide public but may tend to confer some special role of leadership on intellectuals. In this article the noun *ideology* is used only in its strict sense; the adjective *ideological* is used to refer to ideology as broadly defined.

Ideology loosely and strictly defined

On the basis of the five features above, then, one can recognize as ideologies systems as diverse as Destutt de Tracy's own science of ideas, the Positivism of the French philosopher Auguste Comte, Communism and several other types of Socialism, Fascism, Nazism, and certain kinds of nationalism. That all these "-isms" belong to the 19th or 20th century may suggest that ideologies are no older than the word itself—that they belong essentially to a period in which secular belief has increasingly replaced traditional religious faith.

THE PHILOSOPHICAL CONTEXT

Ideology and religion. Ideologies, in fact, are sometimes spoken of as if they belonged to the same logical category as religions. Both are assuredly in a certain sense "total" systems, concerned at the same time with questions of truth and questions of conduct; but the differences between ideologies and religions are perhaps more important than the similarities. A religious theory of reality is constructed in terms of a divine order and is seldom, like that of the ideologist, centred on this world alone. A religion may present a vision of a just society, but it cannot easily have a practical political program. The emphasis of religion is on faith and worship; its appeal is to inwardness and its aim the redemption or purification of the human spirit. An ideology speaks to the group, the nation, or the class. Some religions acknowledge their debt to revelation, whereas ideology always believes, however mistakenly, that it lives by reason alone. Both, it may be said, demand commitment, but it may be doubted whether commitment has ever been a marked feature of those religions into which a believer is inducted in infancy.

Even so, it is in certain religious movements that the first ideological elements in the modern world can be seen. The city of Florence, which in so many fields witnessed

Savonarola as an ideologist

the birth of modernity, produced perhaps the first "ideological" Christian. The attempt of Girolamo Savonarola to construct a puritan utopia was marked by several of the qualities by which one recognizes a modern ideology: Savonarola treated the vision of a Christian community as a model that men should actually seek to realize in the here and now. His method was to dominate the state through an appeal to the populace, and then to use the powers of the state to control both the economy and the private lives of the citizens. The enterprise was given a militant spirit; it was presented by Savonarola as being at one and the same time an outward struggle against papal corruption, the commercial ethos, and Renaissance Humanism, and an inward struggle against worldly ambitions and carnal desires.

Savonarola had numerous followers in his attempt to give Christianity an ideological dimension: he inspired Calvin's Geneva and the Puritan communities of the New World. Indeed, in both the Reformation and the Counter-Reformation, when Christianity was invested with a new militancy and a new intolerance, when a new emphasis was placed on creeds and conversion, religion itself moved that much nearer to ideology.

Ideology in early political philosophy. The Italian political philosopher Niccolò Machiavelli was one of Savonarola's sharpest critics, but he was also, like him, a precursor of modern ideologists. Historians who speak of him only as an immoralist overlook the extent to which Machiavelli was a man with an ideal—a republican ideal. Rousseau recognized this when he spoke of *The Prince* as a "handbook for republicans." Machiavelli's dream was to see revived in modern Italy a republic as glorious as that of ancient Rome, and he suggested that it could be achieved only by means of a revolution that had the strength of will to liquidate its enemies. Machiavelli was the first to link ideology with terror, but he was too much of a political scientist to enact the role of the ideologue.

Ideology in 17th-century England

Seventeenth-century England occupies an important place in the history of ideology. Although there were then no fully fledged ideologies in the strict sense of the term, political theory, like politics itself, began to acquire certain ideological characteristics. The swift movement of revolutionary forces throughout the 17th century created a demand for theories to explain and justify the radical action that was often taken. Locke's *Two Treatises of Government* is an outstanding example of literature written to justify the rights of man against absolutism. This growth of abstract theory in the 17th century, this increasing tendency to construct systems and discuss politics in terms of principles, marks the emergence of the ideological style. In political conversation generally it was accompanied by a growing use of concepts such as right and liberty—ideals in terms of which actual policies were judged.

Hegel and Marx. Although the word *ideology* in the sense derived from Destutt de Tracy's understanding has passed into modern usage, it is important to notice the particular sense that *ideology* is given in Hegelian and Marxist philosophy, where it is used in a pejorative way. *Ideology* there becomes a word for what these philosophers also call "false consciousness." G.W.F. Hegel argued that people were instruments of history; they enacted roles that were assigned to them by forces they did not understand; the meaning of history was hidden from them. Only the philosopher could expect to understand things as they were. This Hegelian enterprise of interpreting reality and reconciling the world to itself was condemned by certain critics as an attempt to provide an ideology of the status quo, in that if individuals were indeed mere ciphers whose actions were determined by external forces, then there was little point in trying to change or improve political and other circumstances. This is a criticism Karl Marx took up, and it is the argument he developed in *The German Ideology* and other earlier writings. Ideology in this sense is a set of beliefs with which people deceive themselves; it is theory that expresses what they are led to think, as opposed to that which is true; it is false consciousness.

Marx, however, was not consistent in his use of the word *ideology*, for he did not always use the term pejoratively, and some of his references to it clearly imply the possibility of an ideology being true. Twentieth-century Marxists, who have frequently discarded the pejorative sense of ideology altogether, have been content to speak of Marxism as being itself an ideology. In certain Communist countries "ideological institutes" have been established, and party philosophers are commonly spoken of as party ideologists. Marxism is an excellent example, a paradigm, of an ideology.

The sociology of knowledge. The use of the word *ideology* in the pejorative sense of false consciousness is found not only in the writings of Marx himself but in those of other exponents of what has come to be known as the sociology of knowledge, including the German sociologists Max Weber and Karl Mannheim, and numerous lesser figures. Few such writers are wholly consistent in their use of the term, but what is characteristic of their approach is their method of regarding idea systems as the outcome or expression of certain interests. In calling such idea systems ideologies, they are treating them as things whose true nature is concealed; they consider the task of sociological research to be the unveiling of what Mannheim called the "life conditions which produce ideologies."

From this perspective, the economic science of Adam Smith, for example, is not to be understood as an independent intellectual construction or to be judged in terms of its truth, consistency, or clarity; rather, it is to be seen as the expression of bourgeois interests, as part of the ideology of capitalism.

Freudian support

The sociology of knowledge in its more recent formulations has sought support in Freudian psychology (notably in borrowing from Freud the concepts of the unconscious and of rationalization), in order to suggest that ideologies are the unconscious rationalizations of class interests. This refinement has enabled sociologists of knowledge to rid their theory of the disagreeable and unscientific element of bald accusation; they no longer have to brand Adam Smith as a deliberate champion of the bourgeois ethos but can see him now as simply the unconscious spokesman of capitalism. At the same time, these sociologists of knowledge have argued that Freudian psychology is itself no less a form of ideology than is Adam Smith's economics, for Freud's method of psychoanalysis is essentially a technique for adjusting rebellious minds to the demands and constraints of bourgeois society.

Critics of the sociology of knowledge have argued that if all philosophy is ideology, then the sociology of knowledge must itself be an ideology like any other idea system and equally devoid of independent validity; that if all seeming truth is veiled rationalization of interest, then the sociology of knowledge cannot be true. It has been suggested that although Weber and Mannheim inspired most of the work that has been done by sociologists of knowledge their own writings may perhaps be exempted from this criticism, if only on the ground that neither of them put forward a consistent or unambiguous theory of ideology. Both used the word *ideology* in different ways at different times. Weber was in part concerned to reverse Marx's theory that all idea systems are products of economic structures, by demonstrating conversely that some economic structures are the product of idea systems (that Protestantism, for example, generated capitalism and not capitalism Protestantism). Mannheim, on the other hand, tried to restore in a more elaborate form Marx's suggestion that ideologies are the product of the social structure. But Mannheim's analysis may have been obscured by his proposal that the word *ideology* should be reserved for idea systems that are more or less conservative, and the word *utopia* for idea systems of a more revolutionary or millenarian nature. Mannheim did not, however, remain faithful to this stipulative definition, even in his book entitled *Ideology and Utopia*.

On the other hand, Mannheim was well aware of the implication of the doctrine that all idea systems have a class basis and a class bias. As a way out of the dilemma he envisaged the possibility of a classless class of intellectuals, a "socially unattached intelligentsia," as he put it, capable of thinking independently by virtue of its independence from any class interest or affiliation. Such a detached group might hope to acquire knowledge that was *not* ideology.

This vision of a small elite of superior minds rising above the myths of ordinary society seemed to some readers to put Mannheim closer to Plato than to Marx and to cast new doubts on the claim of the sociology of knowledge to be a science.

THE POLITICAL CONTEXT

Ideology, rationalism, and romanticism. If some theorists emphasize the kinship between ideology and various forms of religious enthusiasm, others stress the connection between ideology and what they call rationalism, or the attempt to understand politics in terms of abstract ideas rather than of lived experience. Like Napoleon, who held that ideology is par excellence the work of intellectuals, some theorists are suspicious of those who think they know about politics because they have read many books; they believe that politics can be learned only by an apprenticeship to politics itself.

Such people are not unsympathetic to political theories, such as Locke's, but they argue that their value resides in the facts that are derived from experience. Michael Oakeshott in England has described Locke's theory of political liberty as an "abridgment" of the Englishman's traditional understanding of liberty, and has suggested that once such a conception is uprooted from the tradition that has given it meaning it becomes a rationalistic doctrine or metaphysical abstraction, like those liberties contained in the Declaration of the Rights of Man, which were so much talked about after the French Revolution but rarely actually enjoyed, in France or elsewhere.

Whereas Oakeshott has seen ideology as a form of rationalism, Edward Shils, a U.S. political scientist, has seen it more as a product of, among other things, romanticism with an extremist character. His argument is that romanticism has fed into and swelled the seas of ideological politics by its cult of the ideal and by its scorn for the actual, especially its scorn for what is mediated by calculation and compromise. Since civil politics demands both compromise and contrivance and calls for a prudent self-restraint and responsible caution, he suggests that civil politics is bound to be repugnant to romanticism. Hence Shils concludes that the romantic spirit is naturally driven toward ideological politics.

Existential critique of ideology

Ideology and terror. The "total" character of ideology, its extremism and violence, have been analyzed by other critics, among whom the French philosopher-writer Albert Camus and the Austrian-born British philosopher Sir Karl Popper merit particular attention. Beginning as an Existentialist who subscribed to the view that "the universe is absurd," Camus passed to a personal affirmation of justice and human decency as compelling values to be realized in conduct. An Algerian by birth, Camus also appealed to what he believed to be the "Mediterranean" tradition of moderation and human warmth and joy in living as opposed to the "northern" Germanic tradition of fanatical, puritan devotion to metaphysical abstractions. In his book *The Rebel* (*L'Homme révolté*), he argued that the true rebel is not the man who conforms to the orthodoxy of some revolutionary ideology, but a man who could say "no" to injustice. He suggested that the true rebel would prefer the politics of reform, such as that of modern trade-union socialism, to the totalitarian politics of Marxism or similar movements. The systematic violence of ideology—the *crimes de logique* that were committed in its name—appeared to Camus to be wholly unjustifiable. Hating cruelty, he believed that the rise of ideology in the modern world had added enormously to human suffering. Though he was willing to admit that the ultimate aim of most ideologies was to diminish human suffering, he argued that good ends did not authorize the use of evil means.

A somewhat similar plea for what he called "piecemeal social engineering" was put forward by Popper, who argued that ideology rests on a logical mistake: namely the notion that history can be transformed into science. In *The Logic of Scientific Discovery* (*Logik der Forschung*), Popper suggested that the true method of science was not one of observation, hypothesis, and confirmation but one of conjecture and experiment, in which the concept of falsification played a crucial role. By this concept he meant that in science there is a continuing process of trial and error; conjectures are put to the test of experiment, and those that are not falsified are provisionally accepted; thus there is no definitive knowledge but only provisional knowledge that is constantly being corrected. Popper saw in the enterprise of ideology an attempt to find certainty in history and to produce predictions on the model of what were supposed to be scientific predictions. Ideologists, he argued, because they have a false notion of what science is, can produce only prophecies, which are quite distinct from scientific predictions and which have no scientific validity whatever. Though Popper was well disposed toward the idea of a "scientific" approach to politics and ethics, he suggested that a full awareness of the importance of trial and error in science would prompt one to look for similar forms of "negative judgment" elsewhere.

By no means are all ideologists explicit champions of violence, but it is characteristic of ideology both to exalt action and to regard action in terms of a military analogy. Some observers have pointed out that one has only to consider the prose style of the founders of most ideologies to be struck by the military and warlike language that they habitually use, including words like *struggle, resist, march, victory,* and *overcome;* the literature of ideology is replete with martial expressions. In such a view, commitment to an ideology becomes a form of enlistment so that to become the adherent of an ideology is to become a combatant or partisan.

In the years that followed World War II, a number of ideological writers went beyond the mere use of military language and made frank avowals of their desire for violence—not that it was a new thing to praise violence. The French political philosopher Georges Sorel, for example, had done so before World War I in his book *Reflections on Violence*. Sorel was usually regarded as being more a Fascist than a Socialist. He also used the word *violence* in his own special way; by violence Sorel meant passion, not the throwing of bombs and the burning of buildings.

Violence found eloquent champions in several black militant writers of the 1960s, notably the Martinican theorist Frantz Fanon. Moreover, several of the French philosopher Jean-Paul Sartre's dramatic writings turn on the theme that "dirty hands" are necessary in politics and that a man with so-called bourgeois inhibitions about bloodshed cannot usefully serve a revolutionary cause. Sartre's attachment to the ideal of revolution tended to increase as he grew older, and in some of his later writings he suggested that violence might even be a good thing in itself. Sartrian ideology

In considering Sartre's views on the subject of ideology it must be noted that Sartre sometimes used the word *ideology* in a sense peculiarly his own. In an early section of his *Search for a Method* (*Critique de la raison dialectique*), Sartre drew a distinction between philosophies and ideologies in which he reserved the term *philosophy* for those major systems of thought, such as the Rationalism of Descartes or the Idealism of Hegel, which dominate men's minds at a certain moment in history. He defined an ideology as a minor system of ideas, living on the margin of the genuine philosophy and exploiting the domain of the greater system. What Sartre proposed in this work was a revitalization and modernization of the "major philosophy" of Marxism through the integration of elements drawn from the "ideology," or minor system, of Existentialism. What emerged from the book was a theory in which the Existentialist elements are more conspicuous than the Marxist.

Ideology and pragmatism. A distinction is often drawn between the ideological and the pragmatic approach to politics, the latter being understood as the approach that treats particular issues and problems purely on their merits and does not attempt to apply doctrinal, preconceived remedies. Theorists have debated whether or not politics has become less ideological and whether a pragmatic approach can be shown to be better than an ideological one.

On the first question, there seemed to be good reason for thinking that after the death of Stalin and the repudiation of Stalinism by the Communist Party, the Soviet Union, at least, was becoming more interested in the "pragmatic"

concerns of national security and the balance of power and less interested in the ideological aim of fostering universal Communism. This in turn seemed to many to have resulted—in both the United States and the Soviet Union—in a shift toward a pragmatic policy of coexistence and a peaceful division of spheres of influence. There were indications in many countries that the old antagonisms between capitalist and socialist ideologies were giving way to a search for techniques for making a mixed economy work more effectively for the good of all.

But while many observers believed that there was much evidence of a decline of ideology in the latter 1950s, others believed that there were equally manifest signs in the following decade of a revival of ideology, if not within the major political parties, then at least among the public generally. Throughout the world various left-wing movements emerged to challenge the whole ethos on which pragmatic politics was based. Not all these ideologies were coherent, and none possessed the elaborate intellectual structure of the 19th-century ideologies; but together they served to demonstrate that the end of ideology was not yet at hand.

Pragmatism

As suggested earlier, certain controversies about ideology have to some extent been rooted in the ambiguity of the word itself, and this is perhaps especially relevant to the confrontation between ideology and pragmatism, since the word *pragmatism* raises problems no less intractable than those involved in connection with the word *ideology*. In the senses outlined at the beginning of this article, ideology is manifestly not the only alternative to pragmatism in politics, and to reject ideology would not necessarily be to adopt pragmatism. Ordinary language does not yet yield as many words as political science needs to clarify the question, and it becomes necessary to introduce such expressions as belief system, or to name the relevant distinctions, to further the analysis.

Almost any approach to politics constitutes a belief system of one kind or another. Some such belief systems are more structured, more ordered, and generally systematic than others. Though an ideology is a type of belief system, not all belief systems are ideologies. One man's belief system may consist of a congeries of ill-assorted prejudices and inarticulate assumptions. Another's may be the result of deep reflection and careful study. It is sometimes felt to be convenient to speak of a belief system of this latter type as a philosophy or, better, to distinguish it from philosophy in the technical or academic sense, as a *Weltanschauung* (literally, a "view of the world").

The confrontation between ideology and pragmatism may be more instructive if it is translated into a distinction between the ideological and the pragmatic, taking these two adjectives as extremes on a sliding scale. From this perspective, it becomes possible to speak of differences of degree, to speak of an approach to politics as being more or less ideological, more or less pragmatic. At the same time it becomes possible to speak of a belief system such as liberalism as lending itself to a variety of forms, tending at the one extreme toward the ideological, and at the other toward the pragmatic.

THE CONTEXT OF INTERNATIONAL RELATIONS

Dominance of contemporary international relations by ideology

It has been said that ideology has transformed international relationships in the 20th century—in appearance at least. Earlier centuries experienced dynastic wars, national, civil, and imperial wars, and diplomacy designed to further national security or national expansion or to promote mutual advantages and general peace. Such factors, indeed, appeared to govern international relations until recent times. International relations today are seemingly dominated more often than not by the exigencies of "-isms": wars are fought, alliances are made, and treaties are signed because of ideological considerations. The balance of power in the contemporary world is a balance weighted by ideological commitment. "The Communist bloc" confronts "the Free peoples," and in the "Third World" emergent nations cultivate a nationalist, anticolonialist ideology in their search for identity and their efforts to achieve modernity.

But this is not to assert that ideological wars, or ideological diplomacy, are entirely new. What has become the most conspicuous element in contemporary international relations—so conspicuous that other elements are often entirely ignored—was present, to a lesser degree, in earlier international relations. It is necessary here to distinguish between the actual events of history and the interpretations that are put on history, for some events lend themselves more readily than others to an ideological interpretation. The ideological perspective has become increasingly significant as the general public has come to play a role in considering questions of war and peace. When questions of defense and diplomacy were settled by kings and their ministers and wars were fought by professional soldiers and sailors, the public was not expected to have any opinion about international relations, and in such a situation there was little place for ideology.

Ideology in the World Wars. In the course of World War I, however, a new element appeared to have been introduced. The war was seen by those who experienced it as being in its early stages a national war of the traditional kind, and as such it was not at first expected to assume any profoundly disturbing form. Each combatant people viewed itself as fighting for king and country in a just war. But by 1916 the Allies were being urged to think of their endeavour as a war "to make the world safe for democracy," and the Germans, on their side, were correspondingly encouraged to visualize the war as a struggle of "culture" against "barbarism." On both sides, the casualties were far more terrible than anyone had foreseen, and the need to sustain the will to war by an appeal to ideology was plainly felt by all the nations involved. Whether such "war aims" were really the main objectives of the governments concerned is another question; what is important is that, as the need was increasingly felt for a justification of war, the justification took an ideological form. Whether or not World War I changed its real nature between 1914 and 1918, the prevailing conception of it underwent significant alteration. This became more marked after the Russian Revolution of 1917, when the Bolsheviks submitted to harsh German peace terms for reasons that were not only practical but ideological—namely, the preservation and promotion of Communism. Pres. Woodrow Wilson took the United States into the war on the Allied side with an alternative ideological vision—that of ensuring permanent peace through the League of Nations and of establishing democratic governments in all the conquered countries.

Ideology and the will to fight

The rise of Communism clearly marked a corresponding increase in the role of ideology in international relations. Fascism helped to speed the process. The Spanish Civil War of the 1930s was an almost clear-cut confrontation between the ideologies of left and right (not entirely clear-cut because of the ambiguous relationship between Communism and anarchism).

The precise extent of ideological commitment in World War II is a matter of some controversy. At one level, the 1939 war is seen as a continuation of the war of 1914. Two of the leading protagonists—Great Britain and the United States—agreed more in their anti-ideological stance and their hostility to Nazism than in promoting an alternative ideology. Pres. Franklin D. Roosevelt, suspicious of British and French imperialism and eager to cultivate a progressive ideological outlook, was critical of Prime Minister Winston Churchill's politics, hostile toward Charles de Gaulle's, but surprisingly tolerant of Joseph Stalin's. The revival of Wilson's idealistic war aims in the Atlantic Charter provided a basis for a kind of general ideological union of the Allies. But such formulations proved to be of small significance compared to the profound ideological commitment of the Soviet Union to Communism, and that of the United States to an international position more ideologically anti-Communist than pro anything.

"Anti-" and "non-"

Ideology of the Cold War. What came to be called the Cold War in the 1950s must be understood, to a large extent, as an ideological confrontation, and, whereas Communism is manifestly an ideology, the "non-Communism," or even the "anti-Communism," of the West is negatively ideological. To oppose one ideology is not necessarily to subscribe to another, although there is a strong body of opinion in the West that feels that the free world needs a coherent ideology if it is to resist successfully an opposing ideology.

The connection between international wars and ideology can be better expressed in terms of a difference of degree rather than of kind: some wars are more ideological than others, although there is no clear boundary between an ideological and nonideological war. An analogy with the religious wars of the past is evident, and there is indeed some historical continuity between the two types of war. The Christian Crusades against the Turks and the wars between Catholics and Protestants in early modern Europe have much in common with the ideological conflicts of the contemporary period. Religious wars are often communal wars, as witness those between Hindus and Muslims in India; but an "ideological" element of a kind can be discovered in many religious wars, even those narrated in the Old Testament, in which the people of Israel are described as fighting for the cause of righteousness—fighting, in other words, for a universal abstraction as distinct from a local and practical aim. In the past this "ideological" element has in the main been subsidiary; what is characteristic of the modern period is that the ideological element has become increasingly dominant, first in the religious wars (and the related diplomacy) that followed the Reformation and then in the political wars and diplomacy of recent times. (M.C.)

BIBLIOGRAPHY. A useful introduction is M. SELIGER, *Ideology and Politics* (1976), which works from a broad definition of the concept of ideology. JOHN PLAMENATZ, *Ideology* (1970), is a clear and uncomplicated study by a distinguished Oxford philosopher. JEAN BAECHLER, *Qu'est-ce que l'idéologie?* (1976), is characteristically French in its approach and affords an equally lucid introduction to both the sociological and the historical aspects of the problem. Other books written at a fairly popular level include PATRICK CORBETT, *Ideologies* (1966); ROY C. MACRIDIS, *Contemporary Political Ideologies: Movements and Regimes,* 5th ed. (1992); and LEON P. BARADAT, *Political Ideologies: Their Origins and Impact,* 5th ed. (1993).

Few of the works of the original French *idéologues* are available in modern editions and even fewer in English translations. However, RICHARD H. COX (ed.), *Ideology, Politics, and Political Theory* (1969), contains short translated excerpts from Destutt de Tracy and his contemporaries as well as from more recent works. A.L.C. DESTUTT DE TRACY, *A Treatise on Political Economy,* trans. from French, rev. by THOMAS JEFFERSON (1817, reprinted 1973), is his major work in the field; and the expository study by FRANÇOIS JOSEPH PICAVET, *Les Idéologues* (1891, reprinted 1975), remains a classic. The life of Destutt de Tracy and his role in the origins of ideology are traced in EMMET KENNEDY, *A Philosophe in the Age of Revolution: Destutt de Tracy and the Origins of "Ideology"* (1978).

GEORGE LICHTHEIM, *The Concept of Ideology* (1967), contains a short but well-informed and sympathetic analysis of ideology as it figures in Hegelian and Marxist thought. LOUIS ALTHUSSER, *Politics and History: Montesquieu, Rousseau, Hegel, and Marx,* trans. from French (1972, reissued as both *Politics and History: Montesquieu, Rousseau, and Marx* and *Montesquieu, Rousseau, Marx: Politics and History,* 1982), traces the relationship between Hegelian and Marxist thought. G.W.F. HEGEL, *Lectures on the Philosophy of History* (1857, reissued 1956; originally published in German, 3rd ed., 1848), shows relevant elements in his philosophy. Valuable commentaries are provided by ALEXANDRE KOJÈVE, *Introduction to the Reading of Hegel* (1969, reissued 1980; originally published in French, 1947); CHARLES TAYLOR, *Hegel* (1975); and JEAN HYPPOLITE, *Studies on Marx and Hegel* (1969, reissued 1973; originally published in French, 1955). For Marxist philosophy, KARL MARX and FRIEDRICH ENGELS, *The German Ideology,* rev. ed., 2 vol. in 1 (1976; originally published in German, 1932), is the fundamental text. Recent treatments of ideology in the Marxist tradition include ALVIN W. GOULDNER, *The Dialectic of Ideology and Technology* (1976, reissued 1982); JORGE LARRAIN, *The Concept of Ideology* (1979, reprinted 1992), and *Marxism and Ideology* (1983, reprinted 1991); COLIN SUMNER, *Reading Ideologies: An Investigation into the Marxist Theory of Ideology and Law* (1979); and JOE MCCARNEY, *The Real World of Ideology* (1980). More advanced students will find useful WALTER CARLSNAES, *The Concept of Ideology and Political Analysis* (1981).

Writers who have attempted to formulate a neo-Marxist theory of ideology, drawing in part on Hegelian philosophy, include HERBERT MARCUSE, *One Dimensional Man* (1964, reissued 1991); JÜRGEN HABERMAS, *Toward a Rational Society* (1971); and KARL MANNHEIM, *Ideology and Utopia,* new ed. (1991; originally published in German, 1929). Also worthy of attention are LOUIS ALTHUSSER, *Essays on Ideology* (1984); and RAYMOND BOUDON, *The Analysis of Ideology* (1989; originally published in French, 1986).

Interpretations of ideology that are directly opposed to Marxist theory include JAMES R. FLYNN, *Humanism and Ideology* (1973); LEWIS S. FEUER, *Ideology and the Ideologists* (1975); MARTIN SELIGER, *The Marxist Conception of Ideology* (1977); and D.J. MANNING (ed.), *The Form of Ideology* (1980). JEAN-PAUL SARTRE, *Critique of Dialectical Reason* (1976; originally published in French, 1960), constructs a theory of ideology as a "marginal system of ideas" that is consciously designed as an alternative to Marxist theory.

Historical studies that take a relatively extensive view of the impact of ideology as a revolutionary force in the modern world are JAMES H. BILLINGTON, *Fire in the Minds of Men* (1980); MELVIN J. LASKY, *Utopia and Revolution* (1976); and JEANNE HERSCH, *Idéologies et réalité* (1956). HANS KOHN, *Political Ideologies of the Twentieth Century,* 3rd ed. rev. (1966); ISAAC KRAMNICK and FREDERICK M. WATKINS, *The Age of Ideology: Political Thought, 1750 to the Present,* 2nd ed. (1979); and TRYGVE R. THOLFSEN, *Ideology and Revolution in Modern Europe: An Essay on the Role of Ideas in History* (1984), treat ideology as the dominant characteristic of modern political thinking. More polemical commentaries on the development of ideology include ALBERT CAMUS, *The Rebel* (1953, reissued 1991; originally published in French, 1951); JEAN FRANÇOIS REVEL, *Pourquoi des philosophes?* (1957, reissued 1976); and KARL POPPER, *The Poverty of Historicism* (1957, reissued 1986). A systematic critique of the whole notion of ideological politics may be found in MICHAEL OAKESHOTT, *On Human Conduct* (1975, reissued 1991), *On History and Other Essays* (1983), and *Rationalism in Politics,* new and expanded ed. (1991).

RAYMOND ARON, *The Opium of the Intellectuals* (1957, reprinted 1985; originally published in French, 1955), points to a decline in ideological politics in the West; as does DANIEL BELL, *The End of Ideology,* rev. ed. (1962, reissued 1988). Less confident views are advanced in DAVID E. APTER (ed.), *Ideology and Discontent* (1964); and SIDNEY HOOK, *Pragmatism and the Tragic Sense of Life* (1975). An excellent compilation of the contrasting positions in the "End of Ideology" debate is CHAIM I. WAXMAN (ed.), *The End of Ideology Debate* (1968). FRANCIS FUKUYAMA, *The End of History and the Last Man* (1992), asserts that all ideological alternatives to liberal democracy have been discredited.

Sociological aspects of ideology are explored in DONALD G. MACRAE, *Ideology and Society* (1961); NORMAN BIRNBAUM, *The Sociological Study of Ideology (1940–1960)* (1962); ERIC CARLTON, *Ideology and Social Order* (1977); FRANÇOIS BOURRICAUD, *Le Bricolage idéologique* (1980); and GRAHAM C. KNLOCH, *Ideology and Contemporary Sociological Theory* (1981).

The relationship between ideology and political domination is examined in QUINTIN HOARE and GEOFFREY NOWELL SMITH (eds. and trans.), *Selections from the Prison Notebooks of Antonio Gramsci* (1971, reissued 1987). ARNE NAESS, *Democracy, Ideology, and Objectivity* (1956), written from the perspective of political philosophy, was the first of a series of works that investigate the relationship between ideology and liberty. Others worthy of mention are Z.A. JORDAN, *Philosophy and Ideology* (1963); JUDITH N. SHKLAR (ed.), *Political Theory and Ideology* (1966); DANTE GERMINO, *Beyond Ideology* (1967, reprinted 1976); and MAURICE CRANSTON and PETER MAIR (eds.), *Ideology and Politics* (1980). KENNETH MINOGUE, *Alien Powers: The Pure Theory of Ideology* (1985), uses both a philosophical and a historical approach to provide a far-reaching survey of the subject. Among books that stay close to the main tradition of American political science, the following are notable: ROBERT E. LANE, *Political Ideology* (1962); WILLIAM E. CONNOLLY, *Political Science & Ideology* (1967); and ROBERT A. DAHL, *After the Revolution?,* rev. ed. (1990). ANDREW GYORGY and GEORGE D. BLACKWOOD, *Ideologies in World Affairs* (1967), analyzes the emergence of ideology as a decisive factor in international relations. Students interested in such modern ideologies as environmentalism and animal rights should consult IAN ADAMS, *Political Ideology Today* (1993). (M.C./Ed.)

Immunity

Immunity is the ability of humans and other advanced vertebrates to repel disease-causing organisms (pathogens). Immunity from disease is actually conferred by two cooperative mechanisms of the immune system, one called nonspecific, innate immunity and the other specific, acquired immunity. Nonspecific protective mechanisms repel all microorganisms equally, while the specific immune responses are tailored to particular types of invaders. Both systems work together to thwart organisms from entering and proliferating within the body. These immune mechanisms also help eliminate abnormal cells of the body that can develop into cancer.

This article provides a detailed explanation of how nonspecific and specific immunity function and how the immune system evolved. The article also discusses various immune system disorders, including immune deficiencies, allergies, autoimmune disorders, and lymphocyte cancers. For additional information on leukemias, lymphomas, and myelomas, see the article CANCER.

The article is divided into the following sections:

The immune system 773
- Nonspecific, innate immunity 773
  - Barriers against infection
  - Nonspecific responses to infection
- Specific, acquired immunity 775
  - The nature of lymphocytes
  - B-cell antigen receptors and antibodies
  - T-cell antigen receptors
  - Life cycle of lymphocytes
  - Activation of lymphocytes
  - Antibody-mediated immune mechanisms
  - Cell-mediated immune mechanisms
  - Immunity against cancer
  - Prophylactic immunization
  - Production of monoclonal antibodies
- Evolution of the immune system 787
  - Development of immunity in major animal groups
  - Genetic origins of the immune system

Immune system disorders 788
- Immune deficiencies 788
- Allergies 789
  - Type I hypersensitivity
  - Type II hypersensitivity
  - Type III hypersensitivity
  - Type IV hypersensitivity
- Autoimmune disorders 792
  - Basic processes underlying autoimmunity
  - Examples of autoimmune disorders
- Cancers of the lymphocytes 794
  - Genetic causes
  - Malignant transformation of lymphocytes
  - Treatment

Bibliography 794

THE IMMUNE SYSTEM

Nonspecific, innate immunity

BARRIERS AGAINST INFECTION

Most microorganisms encountered in daily life are repelled before they cause detectable signs and symptoms of disease. These potential pathogens, which include viruses, bacteria, fungi, protozoans, and worms, are quite diverse, and therefore a nonspecific defense system that diverts all types of this varied microscopic horde equally is quite useful to an organism. The innate immune system provides this kind of nonspecific protection through a number of defense mechanisms, which include physical barriers such as the skin, chemical barriers such as antimicrobial proteins that harm or destroy invaders, and cells that attack foreign cells and body cells harbouring infectious agents. The details of how these mechanisms operate to protect the body are described in this section.

External barriers. The skin and the mucous membrane linings of the respiratory, gastrointestinal, and genitourinary tracts provide the first line of defense against invasion by microbes or parasites. Human skin has a tough outer layer of cells that produce keratin. This layer of cells, which is constantly renewed from below, serves as a mechanical barrier to infection. In addition, glands in the skin secrete oily substances that include fatty acids, such as oleic acid, that can kill some bacteria; skin glands also secrete lysozyme, an enzyme (also present in tears and saliva) that can break down the outer wall of certain bacteria. Victims of severe burns often fall prey to infections from normally harmless bacteria, illustrating the importance of intact, healthy skin to a healthy immune system.

Like the outer layer of the skin but much softer, the mucous membrane linings of the respiratory, gastrointestinal, and genitourinary tracts provide a mechanical barrier of cells that are constantly being renewed. The lining of the respiratory tract has cells that secrete mucus (phlegm), which traps small particles. Other cells in the wall of the respiratory tract have small hairlike projections called cilia, which steadily beat in a sweeping movement that propels the mucus and any trapped particles up and out of the throat and nose. Also present in the mucus are protective antibodies, which are products of specific immunity. Cells in the lining of the gastrointestinal tract secrete mucus that, in addition to aiding the passage of food, can trap potentially harmful particles or prevent them from attaching to cells that make up the lining of the gut. Protective antibodies are secreted by cells underlying the gastrointestinal lining. Furthermore, the stomach lining secretes hydrochloric acid that is strong enough to kill many microbes.

Chemical barriers. Some microbes penetrate the body's protective barriers and enter the internal tissues. There they encounter a variety of chemical substances that may prevent their growth. These substances include chemicals whose protective effects are incidental to their primary function in the body, chemicals whose principal function is to harm or destroy invaders, and chemicals produced by naturally occurring bacteria.

Some of the chemicals involved in normal body processes are not directly involved in defending the body against disease. Nevertheless, they do help repel invaders. For example, chemicals that inhibit the potentially damaging digestive enzymes released from body cells that have died in the natural course of events also can inhibit similar enzymes produced by bacteria, thereby limiting bacterial growth. Another substance that provides protection against microbes incidentally to its primary cellular role is the blood protein transferrin. The normal function of transferrin is to bind molecules of iron that are absorbed into the bloodstream through the gut and to deliver the iron to cells, which require the mineral to grow. The protective benefit transferrin confers results from the fact that bacteria, like cells, need free iron to grow. As long as it is bound to transferrin, however, iron is unavailable to the invading microbes, and thus their growth is stemmed.

A number of proteins contribute directly to the body's nonspecific defense system by helping to destroy invading microorganisms. One group of such proteins is termed complement because it works with other defense mechanisms of the body, complementing their efforts to eradicate invaders. Many microorganisms can activate complement in ways that do not involve specific immunity. Once activated, complement proteins work together to lyse, or break apart, harmful infectious organisms that do not have protective coats. Other microorganisms can evade these mechanisms but fall prey to scavenger cells, which engulf and destroy infectious agents, and to the mechanisms of the specific immune response. Complement cooperates with both nonspecific and specific defense systems and is described more fully below (see *Antibody-mediated immune mechanisms*).

Interferons

Another group of proteins that provide protection are the interferons, which inhibit the replication of many—but not all—viruses. Cells that have been infected with a virus produce interferon, which sends a signal to other cells of the body to resist viral growth. When first discovered in 1957, interferon was thought to be a single substance, but since then several types have been discovered, each produced by a different type of cell. Alpha interferon is produced by white blood cells other than lymphocytes, beta interferon by fibroblasts, and gamma interferon by lymphocytes. All interferons inhibit viral replication by interfering with the transcription of viral nucleic acid. Interferons exert additional inhibitory effects by regulating the extent to which lymphocytes and other cells express certain important molecules on their surface membranes and by stimulating the activity of natural killer cells, which are described below.

In the small and large intestines the growth of invading bacteria can be inhibited by naturally gut-dwelling bacteria that do not cause disease. These gut-dwelling microorganisms secrete a variety of proteins that enhance their own survival by inhibiting the growth of the invading bacterial species.

Cellular defenses. If an infectious agent is not successfully repelled by the chemical and physical barriers described above, it will encounter cells whose function is to eliminate foreign substances that enter the body. These cells are the nonspecific effector cells of the innate immune response. They include scavenger cells—*i.e.,* various cells that attack infectious agents directly—and natural killer cells, which attack cells of the body that harbour infectious organisms. Some of these cells destroy infectious agents by engulfing and destroying them through the process of phagocytosis, while other cells resort to alternative means. As is true of other components of innate immunity, these cells interact with components of acquired immunity to fight infection.

Scavenger cells. All higher animals and many lower ones have scavenger cells—primarily leukocytes (white blood cells)—that destroy infectious agents. Most vertebrates, including all birds and mammals, possess two main kinds of scavenger cells. Their importance was first recognized in 1884 by the Russian biologist Élie Metchnikoff, who named them microphages and macrophages, after Greek words meaning "little eaters" and "big eaters."

Granulocytes

Microphages are now called either granulocytes, because of the numerous chemical-containing granules found in their cytoplasm, or polymorphonuclear leukocytes, because of the oddly shaped nucleus these cells contain. Some granules contain digestive enzymes capable of breaking down proteins, while others contain bacteriocidal (bacteria-killing) proteins. There are three classes of granulocytes—neutrophils, eosinophils, and basophils—that are distinguished according to the shape of the nucleus and the way in which the granules in the cytoplasm are stained by dye. The differences in staining characteristics reflect differences in the chemical makeup of the granules. Neutrophils are the most common type of granulocyte, making up about 60 to 70 percent of all white blood cells. These granulocytes ingest and destroy microorganisms, especially bacteria. Less common are the eosinophils, which are particularly effective at damaging the cells that make up the cuticle (body wall) of larger parasites. Fewer still are the basophils, which release heparin (a substance that inhibits blood coagulation), histamine, and other substances that play a role in some allergic reactions (see below *Immune system disorders: Allergies*). Very similar in structure and function to basophils are the tissue cells called mast cells, which also contribute to immune responses.

Granulocytes, which have a life span of only a few days, are continuously produced from stem (*i.e.,* precursor) cells in the bone marrow. They enter the bloodstream and circulate for a few hours, after which they leave the circulation and die. Granulocytes are mobile and are attracted to foreign materials by chemical signals, some of which are produced by the invading microorganisms themselves, others by damaged tissues, and still others by the interaction between microbes and proteins in the blood plasma. Some microorganisms produce toxins that poison granulocytes and thus escape phagocytosis; other microbes are indigestible and are not killed when ingested. By themselves, then, granulocytes are of limited effectiveness and require reinforcement by the mechanisms of specific immunity.

Macrophages

The other main type of scavenger cell is the macrophage, the mature form of the monocyte (see Figure 1). Like granulocytes, monocytes are produced by stem cells in the bone marrow and circulate through the blood, though in lesser numbers. But, unlike granulocytes, monocytes undergo differentiation, becoming macrophages that settle in many tissues, especially the lymphoid tissues (*e.g.,* spleen and lymph nodes) and the liver, which serve as filters for trapping microbes and other foreign particles that arrive through the blood or the lymph. Macrophages live longer than granulocytes and, although effective as scavengers, basically provide a different function. Compared with granulocytes, macrophages move relatively sluggishly. They are attracted by different stimuli and usually arrive at sites of invasion later than granulocytes. Macrophages recognize and ingest foreign particles by mechanisms that are basically similar to those of granulocytes, although the digestive process is slower and not as complete. This aspect is of great importance for the role that macrophages play in stimulating specific immune responses—something in which granulocytes play no part (see below *Activation of lymphocytes*).

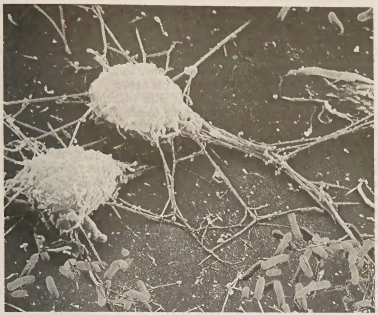
Natural killer (NK) cells. Natural killer cells do not attack invading organisms directly but instead destroy the body's own cells that have either become cancerous or been infected with a virus. NK cells were first recognized in 1975, when researchers observed cells in the blood and lymphoid tissues that were neither the scavengers described above nor ordinary lymphocytes but which nevertheless were capable of killing cells. Although similar in outward appearance to lymphocytes, NK cells contain granules that harbour cytotoxic chemicals. NK cells recognize dividing cells by a mechanism that does not depend on specific immunity. They then bind to these dividing cells and insert their granules through the outer membrane and into the cytoplasm. This causes the dividing cells to leak and die. It is not certain whether NK cells belong to a distinct lineage or are a special form of lymphocyte. It is known that they are stimulated by gamma interferon. Their main biological role may be to regulate the growth of stem cells in the bone marrow and elsewhere.

NONSPECIFIC RESPONSES TO INFECTION

The body has a number of nonspecific methods of fighting infection that are called early induced responses. They include the acute-phase response and the inflammation response, which can eliminate infection or hold it in check until specific, acquired immune responses have time to develop. Nonspecific immune responses occur more rapidly than acquired immune responses do, but they do not provide lasting immunity to specific pathogens.

Nonadaptive immune responses rely on a number of chemical signals, collectively called cytokines, to carry out their effects. These cytokines include members of the family of proteins called interleukins, which induce fever and the acute-phase response, and tumour necrosis factor-alpha, which initiates the inflammatory response.

Acute-phase response. When the body is invaded by a pathogen, macrophages release the protein signals inter-



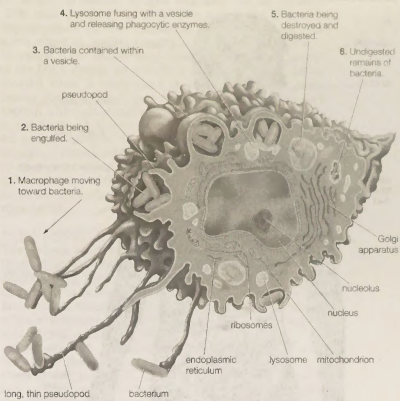


Figure 1: *The destruction of bacteria by a macrophage, one of the principal phagocytic (cell-engulfing) components of the immune system.*
(Left) Electron micrograph of macrophages attacking *Escherichia coli*. (Right) Drawing of the internal and external structures of a macrophage.
(Left) © Manfred Kage/Peter Arnold, Inc., (right) Encyclopaedia Britannica, Inc.

leukin-1 (IL-1) and interleukin-6 (IL-6) to help fight the infection. One of their effects is to raise the temperature of the body, causing the fever that often accompanies infection. (The interleukins increase body temperature by acting on the temperature-regulating hypothalamus in the brain and by affecting energy mobilization by fat and muscle cells.) Fever is believed to be helpful in eliminating infections because most bacteria grow optimally at temperatures lower than normal body temperature. But fever is only part of the more general innate defense mechanism called the acute-phase response. In addition to raising body temperature, the interleukins stimulate liver cells to secrete increased amounts of several different proteins into the bloodstream. These proteins, collectively called acute-phase proteins, bind to bacteria and, by doing so, activate complement proteins that destroy the pathogen. The acute-phase proteins act similarly to antibodies but are more democratic—that is, they do not distinguish between pathogens as antibodies do but instead attack a wide range of microorganisms equally. Another effect the interleukins have is to increase the number of circulating neutrophils and eosinophils, which help fight infection.

Interleukins

Inflammatory response. Infection often results in tissue damage, which may trigger an inflammatory response. The signs of inflammation include pain, swelling, redness, and fever, which are induced by chemicals released by macrophages. These substances promote blood flow to the area, increase the permeability of capillaries, and induce coagulation. The increased blood flow is responsible for redness, and the leakiness of the capillaries allows cells and fluids to enter tissues, causing pain and swelling. These effects bring more phagocytic cells to the area to help eliminate the pathogens. The first cells to arrive, usually within an hour, are neutrophils and eosinophils, followed a few hours later by macrophages. Macrophages not only engulf pathogens but also help the healing process by disposing of cellular debris which accumulates from destroyed tissue cells and neutrophils that self-destruct after ingesting microorganisms. If infection persists, components of specific immunity—antibodies and T cells—arrive at the site to fight the infection.

Specific, acquired immunity

It has been known for centuries that persons who have contracted certain diseases and survived generally do not catch those illnesses again. The Greek historian Thucydides recorded that, when the plague was raging in Athens during the 5th century BC, the sick and dying would have received no nursing at all had it not been for the devotion of those who had already recovered from the disease; it was known that no one ever caught the plague a second time. The same applies, with rare exceptions, to many other diseases, such as smallpox, chicken pox, measles, and mumps. Yet having had measles does not prevent a child from contracting chicken pox, or vice versa. The protection acquired by experiencing one of these infections is specific for that infection; in other words, it is due to specific, acquired immunity, also called adaptive immunity.

There are other infectious conditions, such as the common cold, influenza, pneumonia, and diarrheal diseases, which can be caught again and again; these seem to contradict the notion of specific immunity. But the reason such illnesses can recur is that many different infectious agents produce similar symptoms (and thus the same disease). For example, more than 100 viruses can cause the cluster of symptoms known as the common cold. Consequently, even though infection with a particular agent does protect against reinfection by that same pathogen, it does not confer protection from other pathogens that have not been encountered.

Acquired immunity is dependent on the specialized white blood cells known as lymphocytes. This section describes the various ways in which lymphocytes operate to confer specific immunity.

THE NATURE OF LYMPHOCYTES

General characteristics. *Location in the immune system.* Lymphocytes are the cells responsible for the body's ability to distinguish and react to an almost infinite number of different foreign substances, including those of which microbes are composed. Lymphocytes are mainly a dormant population, awaiting the appropriate signals to be stirred to action. The inactive lymphocytes are small, round cells filled largely by a nucleus. Although they have only a small amount of cytoplasm compared with other cells, each lymphocyte has sufficient cytoplasmic organelles (small functional units such as mitochondria, the endoplasmic reticulum, and a Golgi apparatus) to keep the cell alive. Lymphocytes move only sluggishly on their own, but they can travel swiftly around the body when carried along in the blood or lymph. At any one time an adult human has approximately 2×10^{12} lymphocytes, about 1 percent of which are in the bloodstream. The majority are concentrated in various tissues scattered throughout the body, particularly the bone marrow, spleen, thymus, lymph nodes, tonsils, and lining of the intestines, which make up the lymphatic system (see Figure 2). Organs or tissues containing such concentrations of lymphocytes are termed

Lymphoid tissues

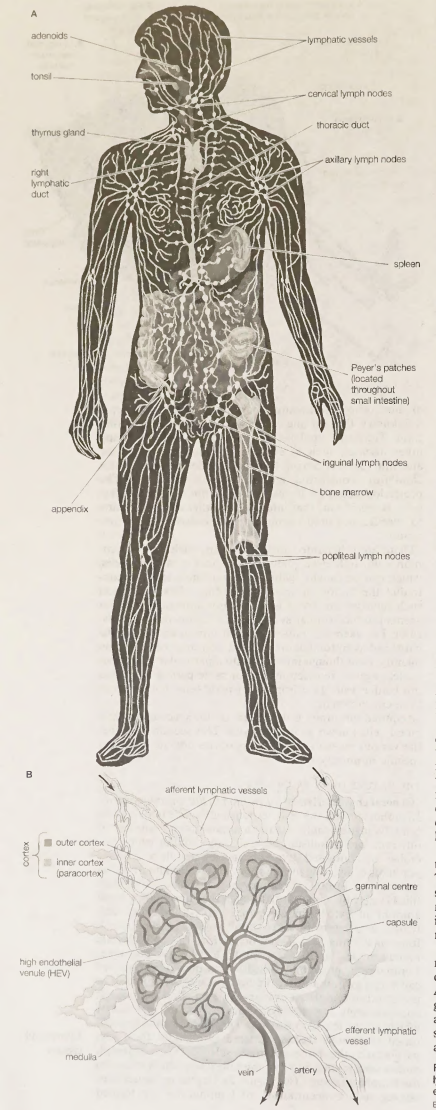


Figure 2: (A) The human lymphatic system, showing the lymphatic vessels and lymphoid organs. (B) Internal and external structures of a lymph node.

Encyclopædia Britannica, Inc.

lymphoid. The lymphocytes in lymphoid structures are free to move, although they are not lying loose; rather, they are confined within a delicate network of lymph capillaries located in connective tissues that channel the lymphocytes so that they come into contact with other cells, especially macrophages, that line the meshes of the network. This ensures that the lymphocytes interact with each other and with foreign materials trapped by the macrophages in an ordered manner.

T and B cells. Lymphocytes originate from stem cells in the bone marrow; these stem cells divide continuously, releasing immature lymphocytes into the bloodstream. Some of these cells travel to the thymus, where they multiply and differentiate into T lymphocytes, or T cells. The "T" stands for *t*hymus-derived, referring to the fact that these cells mature in the thymus. Once they have left the thymus, T cells enter the bloodstream and circulate to and within the rest of the lymphoid organs, where they can multiply further in response to appropriate stimulation. About half of all lymphocytes are T cells.

Some lymphocytes remain in the bone marrow, where they differentiate and then pass directly to the lymphoid organs. They are termed B lymphocytes, or B cells, and they, like T cells, can mature and multiply further in the lymphoid organs when suitably stimulated. Although it is appropriate to refer to them as B cells in humans and other mammals, because they are bone-marrow derived, the "B" actually stands for the *b*ursa of Fabricius, a lymphoid organ found only in birds, the organisms in which B cells were first discovered.

B and T cells both recognize and help eliminate foreign molecules (antigens), such as those that are part of invading organisms, but they do so in different ways. B cells secrete antibodies, proteins that bind to antigens. Since antibodies circulate through the humours (*i.e.,* body fluids), the protection afforded by B cells is called humoral immunity. T cells, in contrast, do not produce antibodies but instead directly attack invaders. Because this second type of acquired immunity depends on the direct involvement of cells rather than antibodies, it is called cell-mediated immunity. T cells recognize only infectious agents that have entered into cells of the body, whereas B cells and antibodies interact with invaders that remain outside the body's cells. These two types of specific, acquired immunity, however, are not as distinct as might be inferred from this description, since T cells also play a major role in regulating the function of B cells. In many cases an immune response involves both humoral and cell-mediated assaults on the foreign substance. Furthermore, both classes of lymphocytes can activate or enhance a variety of nonspecific immune responses.

Ability to recognize foreign molecules. *Receptor molecules.* Lymphocytes are distinguished from other cells by their capacity to recognize foreign molecules. Recognition is accomplished by means of receptor molecules. A receptor molecule is a special protein whose shape is complementary to a portion of a foreign molecule. This complementarity of shape allows the receptor and the foreign molecule to conform to each other in a fashion roughly analogous to the way a key fits into a lock.

Receptor molecules are either attached to the surface of the lymphocyte or secreted into fluids of the body. B and T lymphocytes both have receptor molecules on their cell surfaces, but only B cells manufacture and secrete large numbers of unattached receptor molecules, called antibodies. Antibodies correspond in structure to the receptor molecules on the surface of the B cell.

Antigens. Any foreign material—usually of a complex nature and often a protein—that binds specifically to a receptor molecule made by lymphocytes is called an antigen. Antigens include molecules found on invading microorganisms, such as viruses, bacteria, protozoans, and fungi, as well as molecules located on the surface of foreign substances, such as pollen, dust, or transplanted tissue. When an antigen binds to a receptor molecule, it may or may not

Antigens and immune response

evoke an immune response. Antigens that induce such a response are called immunogens. Thus, it can be said that all immunogens are antigens, but not all antigens are immunogens. For example, a simple chemical group that can combine with a lymphocyte receptor (*i.e.*, is an antigen) but does not induce an immune response (*i.e.*, is not an immunogen) is called a hapten. Although haptens cannot evoke an immune response by themselves, they can become immunogenic when joined to a larger, more complex molecule such as a protein, a feature that is useful in the study of immune responses.

Many antigens have a variety of distinct three-dimensional patterns on different areas of their surfaces. Each pattern is called an antigenic determinant, or epitope, and each epitope is capable of reacting with a different lymphocyte receptor. Complex antigens present an "antigenic mosaic" and can evoke responses from a variety of specific lymphocytes. Some antigenic determinants are better than others at effecting an immune response, presumably because a greater number of responsive lymphocytes are present. It is possible for two or more different substances to have an epitope in common. In these cases, immune components induced by one antigen are able to react with all other antigens carrying the same epitope. Such antigens are known as cross-reacting antigens.

T cells and B cells differ in the form of the antigen they recognize, and this affects which antigens they can detect. B cells bind to antigen on invaders that are found in circulation outside the cells of the body, while T cells detect only invaders that have somehow entered the cells of the body. Thus foreign materials that have been ingested by cells of the body or microorganisms such as viruses that penetrate cells and multiply within them are out of reach of antibodies but can be eliminated by T cells.

Diversity of lymphocytes. The specific immune system (in other words, the sum total of all the lymphocytes) can recognize virtually any complex molecule that nature or science has devised. This remarkable ability results from the trillions of different antigen receptors that are produced by the B and T lymphocytes. Each lymphocyte produces its own specific receptor, which is structurally organized so that it responds to a different antigen. After a cell encounters an antigen that it recognizes, it is stimulated to multiply, and the population of lymphocytes bearing that particular receptor increases.

How is it that the body has such an incredible diversity of receptors that are always ready to respond to invading molecules? To understand this, a quick review of genes and proteins will be helpful. Antigen receptor molecules are proteins, which are composed of a few polypeptide chains (*i.e.*, chains of amino acids linked together by chemical bonds known as peptide bonds). The sequence in which the amino acids are assembled to form a particular polypeptide chain is specified by a discrete region of DNA called a gene. But, if every polypeptide region of every antigen receptor were encoded by a different gene, the human genome (all the genetic information encoded in the DNA that is carried on the chromosomes of cells) would need to devote trillions of genes to code just for these immune system proteins. Since the entire human genome contains approximately 30,000 genes, individuals cannot inherit a gene for each particular antigen receptor component. Instead, a mechanism exists that generates an enormous variety of receptors from a limited number of genes.

Genetic coding of antigen receptors

What is inherited is a pool of gene segments for each type of polypeptide chain. As each lymphocyte matures, these gene segments are pieced together to form one gene for each polypeptide that makes up a specific antigen receptor. This rearrangement of alternative gene segments occurs predominantly, though not entirely, at random, so that an enormous number of combinations can result. Additional diversity is generated from the imprecise recombination of gene segments—a process called junctional diversification—through which the ends of the gene segments can be shortened or lengthened. The genetic rearrangement takes place at the stage when the lymphocytes generated from stem cells first become functional, so that each mature lymphocyte is able to make only one type of receptor. Thus, from a pool of only hundreds of genes, an unlimited variety of diverse antigen receptors can be created.

Still other mechanisms contribute to receptor diversity. Superimposed on the mechanism outlined in simplified terms above is another process, called somatic mutation. Mutation is the spontaneous occurrence of small changes in the DNA during the process of cell division. It is called somatic when it takes place in body cells (Greek *soma* means "body") rather than in germ-line cells (eggs and sperm). Although somatic mutation can be a chance event in any body cell, it occurs regularly in the DNA that codes for antigen receptors in lymphocytes. Thus, when a lymphocyte is stimulated by an antigen to divide, new variants of its antigen receptor can be present on its descendant cells, and some of these variants may provide an even better fit for the antigen that was responsible for the original stimulation.

B-CELL ANTIGEN RECEPTORS AND ANTIBODIES

The antigen receptors on B lymphocytes are identical to the binding sites of antibodies that these lymphocytes manufacture once stimulated, except that the receptor molecules have an extra tail that penetrates the cell membrane and anchors them to the cell surface. Thus, a description of the structure and properties of antibodies, which are well studied, will suffice for both.

Basic structure. Antibodies belong to the class of proteins called globulins, so named for their globular structure. Collectively, antibodies are known as immunoglobulins (abbreviated Ig). All immunoglobulins have the same basic molecular structure, consisting of four polypeptide chains (see Figure 3). Two of the chains, which are identical in any given immunoglobulin molecule, are heavy (H) chains; the other two are identical light (L) chains. The terms "heavy" and "light" simply mean larger and smaller. Each chain is manufactured separately and is encoded by different genes. The four chains are joined in the final immunoglobulin molecule to form a flexible Y shape, which is the simplest form an antibody can take.

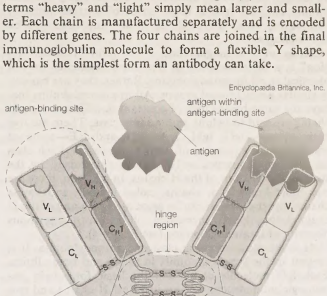


Figure 3: *The four-chain structure of an antibody, or immunoglobulin, molecule.*
The basic unit is composed of two identical light (L) chains and two identical heavy (H) chains, which are held together by disulfide bonds to form a flexible Y shape. Each chain is composed of a variable (V) region and a constant (C) region.

Binding of antibody to antigen

At the tip of each arm of the Y-shaped molecule is an area called the antigen-binding, or antibody-combining, site, which is formed by a portion of the heavy and light chains. Every immunoglobulin molecule has at least two of these sites, which are identical to one another. The antigen-binding site is what allows the antibody to recognize a specific part of the antigen (the epitope, or antigenic determinant). If the shape of the epitope corresponds to the shape of the antigen-binding site, it can fit into the site—that is, be "rec-

ognized" by the antibody. Chemical bonds called weak bonds then form to hold the antigen within the binding site.

The heavy and light chains that make up each arm of the antibody are composed of two regions, called constant (C) and variable (V). These regions are distinguished on the basis of amino acid similarity—that is, constant regions have essentially the same amino acid sequence in all antibody molecules of the same class (IgG, IgM, IgA, IgD, or IgE), but the amino acid sequences of the variable regions differ quite a lot from antibody to antibody. This makes sense, because the variable regions determine the unique shape of the antibody-binding site. The tail of the molecule, which does not bind to antigens, is composed entirely of the constant regions of heavy chains.

The variable and constant regions of both the light and the heavy chains are structurally folded into functional units called domains. Each light chain consists of one variable domain (V_L) and one constant domain (C_L). Each heavy chain has one variable domain (V_H) and three or four constant domains (C_H1, C_H2, C_H3, C_H4). Those domains that make up the "tail" of the basic Y-shaped molecule (in other words, all the H-chain constant domains except C_H1) are responsible for the special biological properties of immunoglobulins—except, of course, for the capacity to bind to a specific antigenic determinant. The tail of the antibody determines the fate of the antigen once it becomes bound to the antibody.

The hinge region of the antibody is a short stretch of amino acids on the heavy chain located between the chain's C_H1 and C_H2 regions. It provides the molecule with flexibility, which is very useful in binding antigens. This flexibility can actually improve the efficiency with which an antigen binds to the antibody. It can also help in cross-linking antigens into a large lattice of antigen-antibody complexes, which are easily identified and destroyed by macrophages.

Classes of immunoglobulins. The term "constant region" is a bit misleading in that these segments are not identical in all immunoglobulins. Rather, they are basically similar among broad groups. All immunoglobulins that have the same basic kinds of constant domains in their H chains are said to belong to the same class. There are five main classes—IgG, IgM, IgA, IgD, and IgE—some of which include a number of distinct subclasses. Each class has its own properties and functions determined by the structural variations of the H chains. In addition, there are two basic kinds of L chains, called lambda and kappa chains, either of which can be associated with any of the H chain classes, thereby increasing still further the enormous diversity of immunoglobulins (see Figure 4).

IgG is the most common class of immunoglobulin. It is present in the largest amounts in blood and tissue fluids. Each IgG molecule consists of the basic four-chain immunoglobulin structure—two identical H chains and two identical L chains (either kappa or lambda)—and thus carries two identical antigen-binding sites. There are four subclasses of IgG, each with minor differences in its H chains but with distinct biological properties. IgG is the only class of immunoglobulin capable of crossing the placenta; consequently, it provides some degree of immune protection to the developing fetus. These molecules also are secreted into the mother's milk and, once they have been ingested by an infant, can be transported into the blood, where they confer immunity.

IgM is the first class of immunoglobulin made by B cells as they mature, and it is the form most commonly present as the antigen receptor on the B-cell surface. When IgM is secreted from the cells, five of the basic Y-shaped units become joined together to make a large pentamer molecule with 10 antigen-binding sites. This large antibody molecule is particularly effective at attaching to antigenic determinants present on the outer coats of bacteria. When this IgM attachment occurs, it causes microorganisms to agglutinate, or clump together.

IgA is the main class of antibody found in many body secretions, including tears, saliva, respiratory and intestinal secretions, and colostrum (the first milk produced by lactating mothers). Very little IgA is present in the serum. IgA is produced by B cells located in the mucous membranes

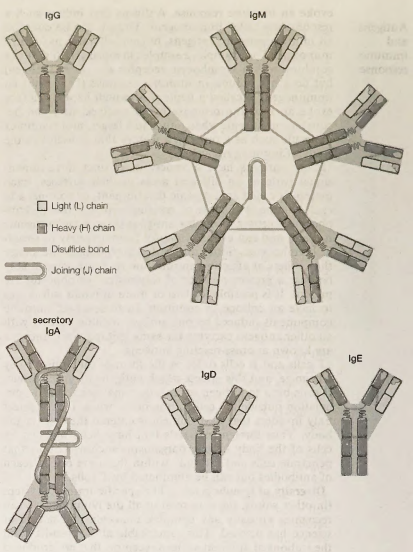


Figure 4: The five main classes of antibodies (immunoglobulins): IgG, IgM, IgA, IgD, and IgE.
Encyclopædia Britannica, Inc.

of the body. Two molecules of IgA are joined together and associated with a special protein that enables the newly formed IgA molecule to be secreted across epithelial cells that line various ducts and organs. Although IgG is the most common class of immunoglobulin, more IgA is synthesized by the body daily than any other class of antibody. However, IgA is not as stable as IgG, and therefore it is present in lower amounts at any given time.

IgD molecules are present on the surface of most, but not all, B cells early in their development, but little IgD is ever released into the circulation. It is not clear what function IgD performs, though it may play a role in determining whether antigens activate the B cells.

IgE is made by a small proportion of B cells and is present in the blood in low concentrations. Each molecule of IgE consists of one four-chain unit and so has two antigen-binding sites, like the IgG molecule; however, each of its H chains has an extra constant domain (C_H4), which confers on IgE the special property of binding to the surface of basophils and mast cells. When antigens bind to these attached IgE molecules, the cell is stimulated to release chemicals, such as histamines, that are involved in allergic reactions (see below *Immune system disorders: Type I hypersensitivity*). IgE antibodies also help to protect against parasitic infections.

Normal production of antibody. Most individuals have fairly constant amounts of immunoglobulin in their blood, which represent the balance between continuous breakdown of these proteins and their manufacture. There is about 4 times as much IgG (including its subclasses) as IgA, 10 to 15 times as much as IgM, 300 times as much as IgD, and 30,000 times as much as IgE.

Part of the normal production of immunoglobulin undoubtedly represents the response to antigenic stimulation that happens continually, but even animals raised in surroundings completely free from microbes and their products make substantial, though lesser, amounts of

immunoglobulin. Much of the immunoglobulin therefore must represent the product of all the B cells that are, so to speak, "ticking over" even if not specifically stimulated. It is therefore not surprising that extremely sensitive methods can detect traces of antibodies that react with antigenic determinants to which an animal has never been exposed but for which cells with receptors are present.

All B cells have the potential to use any one of the constant-region classes to make up the immunoglobulin they secrete. As noted above, when first stimulated, most secrete IgM. Some continue to do so, but others later switch to producing IgG, IgA, or IgE. Memory B cells, which are specialized for responding to repeat infections by a given antigen, make IgG or IgA immediately (see below *Activation of lymphocytes*). What determines the balance among the classes of antibodies is not fully understood. However, it is influenced by the nature and site of deposition of the antigen (for example, parasites tend to elicit IgE), and their production is clearly mediated by factors, called cytokines, which are released locally by T cells.

Memory cells and long-term immunity

T-CELL ANTIGEN RECEPTORS

Structure. T-cell antigen receptors are found only on the cell membrane. For this reason, T-cell receptors were difficult to isolate in the laboratory and were not identified until 1983. T-cell receptors consist of two polypeptide chains. The most common type of receptor is called alpha-beta because it is composed of two different chains, one called alpha and the other beta (see Figure 5). A less common type is the gamma-delta receptor, which contains a different set of chains, one gamma and one delta. A typical T cell may have as many as 20,000 receptor molecules on its membrane surface, all of either the alpha-beta or gamma-delta type.

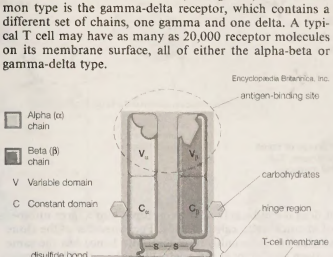


Figure 5: The basic structure of a typical T-cell receptor.

The T-cell receptor molecule is embedded in the membrane of the cell, and a portion of the molecule extends away from the cell surface into the area surrounding the cell. The chains each contain two folded domains, one constant and one variable, an arrangement similar to that of the chains of antibody molecules. And, as is true of antibody structure, the variable domains of the chains form an antigen-binding site. However, the T-cell receptor has only one antigen-binding site, unlike the basic antibody molecule, which has two.

Many structural similarities exist between antibodies and T-cell receptors. Therefore, it is not surprising that the organization of genes that encode the T-cell receptor chains is similar to that of immunoglobulin genes. Similarities also exist between the mechanisms B cells use to generate antibody diversity and those used by T cells to create T-cell diversity. These commonalities suggest that both systems evolved from a more primitive and simpler recognition system (see below *Genetic origins of the immune system*).

Function. Despite the structural similarities, the receptors on T cells function differently from those on B cells. The functional difference underlies the different roles played by B and T cells in the immune system. B cells secrete antibodies to antigens in blood and other body fluids, but T cells cannot bind to free-floating antigens. Instead they bind to fragments of foreign proteins that are displayed on the surface of body cells. Thus, once a virus succeeds in infecting a cell, it is removed from the reach of circulating antibodies only to become susceptible to the defense system of the T cell.

But how do fragments of a foreign substance come to be displayed on the surface of a body cell? First, as is shown in Figure 6, the substance must enter the cell, which can happen through either phagocytosis or infection. Next, the invader is partially digested by the body cell, and one of its fragments is moved to the surface of the cell, where it becomes bound to a cell-surface protein. This cell-surface protein is the product of one of a group of molecules encoded by the genes of the major histocompatibility complex (MHC). In humans MHC proteins were first discovered on leukocytes (white blood cells) and, therefore, are often referred to as HLA (human leukocyte antigens). There are two major types of MHC molecules: class I molecules, which are present on the surfaces of virtually all cells of the body that contain nuclei—that is, most body cells—and class II molecules, which are restricted to the surfaces of most B cells and some T cells, macrophages, and macrophage-like cells.

Major histocompatibility complex

Two main types of mature T cells—cytotoxic T cells and helper T cells—are known. Some scientists hypothesize the existence of a third type of mature T cell called regulatory T cells. Some T cells recognize class I MHC molecules on the surface of cells; others bind to class II molecules. Cytotoxic T cells destroy body cells that pose a threat to the individual—namely, cancer cells and cells containing harmful microorganisms. Helper T cells do not directly kill other cells but instead help activate other white blood cells (lymphocytes and macrophages), primarily by secreting a variety of cytokines that mediate changes in other cells. The function of regulatory T cells is poorly understood. To carry out their roles, helper T cells recognize foreign antigens in association with class II MHC molecules on the surfaces of macrophages or B cells. Cytotoxic T cells and regulatory T cells generally recognize target cells bearing antigens associated with class I molecules. Because they recognize the same class of MHC molecule, cytotoxic and regulatory T cells are often grouped together; however, populations of both types of cells associated with class II molecules have been reported. Cytotoxic T cells can bind to virtually any cell in the body that has been invaded by a pathogen.

T cells have another receptor, or coreceptor, on their surface that binds to the MHC molecule and provides additional strength to the bond between the T cell and the target cell. Helper T cells display a coreceptor called CD4, which binds to class II MHC molecules, and cytotoxic T cells have on their surfaces the coreceptor CD8, which recognizes class I MHC molecules. These accessory receptors add strength to the bond between the T cell and the target cell.

The T-cell receptor is associated with a group of molecules called the CD3 complex, or simply CD3, which is also necessary for T-cell activation. These molecules are agents that help transduce, or convert, the extracellular binding of the antigen and receptor into internal cellular signals; thus, they are called signal transducers. Similar signal transducing molecules are associated with B-cell receptors.

LIFE CYCLE OF LYMPHOCYTES

T cells. When T-cell precursors leave the bone marrow on their way to mature in the thymus, they do not yet express receptors for antigens and thus are indifferent to stimulation by them. Within the thymus the T cells multiply many times as they pass through a meshwork of thymus cells. In the course of multiplication they acquire antigen receptors and differentiate into helper or cytotoxic T cells. As mentioned in the previous section, these cell types, similar in appearance, can be distinguished by their function and by the presence of the special surface proteins, CD4 and CD8. Most T cells that multiply in the thymus also die there. This seems wasteful until it is remembered that the random generation of different antigen receptors yields a large proportion of receptors that

Helper T cells and cytotoxic T cells

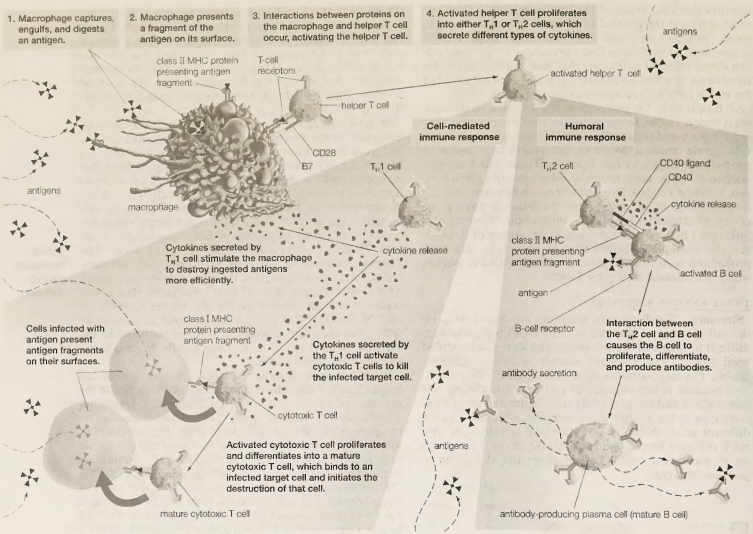


Figure 6: *Stimulation of immune response by activated helper T cells.* Activated by complex interaction with molecules on the surface of a macrophage or some other antigen-presenting cell, a helper T cell proliferates into two general subtypes, T_H1 and T_H2. These in turn stimulate the complex pathways of the cell-mediated immune response and the humoral immune response, respectively.
Encyclopædia Britannica, Inc.

recognize self antigens—*i.e.,* molecules present on the body's own constituents—and that mature lymphocytes with such receptors would attack the body's own tissues. Most such self-reactive T cells die before they leave the thymus, so that those T cells that do emerge are the ones capable of recognizing foreign antigens. These travel via the blood to the lymphoid tissues, where, if suitably stimulated, they can again multiply and take part in immune reactions. The generation of T cells in the thymus is an ongoing process in young animals. In humans large numbers of T cells are produced before birth, but production gradually slows down during adulthood and is much diminished in old age, by which time the thymus has become small and partly atrophied. Cell-mediated immunity persists throughout life, however, because some of the T cells that have emerged from the thymus continue to divide and function for a very long time.

B cells. B-cell precursors are continuously generated in the bone marrow throughout life, but, as with T-cell generation, the rate diminishes with age. Unless they are stimulated to mature (as described below), the majority of B cells also die, although those that have matured can survive for a long time in the lymphoid tissues. Consequently, there is a continuous supply of new B cells throughout life. Those with antigen receptors capable of recognizing self antigens tend to be eliminated, though less effectively than are self-reactive T cells. As a result, some self-reactive cells are always present in the B-cell population, along with the majority that recognize foreign antigens. The reason the self-reactive B cells normally do no harm is explained in the following section.

ACTIVATION OF LYMPHOCYTES

In its lifetime a lymphocyte may or may not come into contact with the antigen it is capable of recognizing, but if it does it can be activated to multiply into a large number of identical cells, called a clone. Each member of the clone carries the same antigen receptor and hence has the same antigen specificity as the original lymphocyte. The process, called clonal selection, is one of the fundamental concepts of immunology.

Two types of cells are produced by clonal selection—effector cells and memory cells. Effector cells are the relatively short-lived activated cells that defend the body in an immune response. Effector B cells are called plasma cells and secrete antibodies, and activated T cells include cytotoxic T cells and helper T cells, which carry out cell-mediated responses. The production of effector cells in response to first-time exposure to an antigen is called the primary immune response. Memory cells also are produced at this time, but they do not become active at this point. However, if the organism is reexposed to the same antigen that stimulated their formation, the body mounts a second immune response that is led by these long-lasting memory cells, which then give rise to another population of identical effector and memory cells. This secondary mechanism is known as immunological memory, and it is responsible for the lifetime immunities to diseases such as measles that arise from childhood exposure to the causative pathogen.

T cells. Helper T cells do not directly kill infected cells, as cytotoxic T cells do. Instead they help activate cytotoxic T cells and macrophages to attack infected cells, or they stimulate B cells to secrete antibodies. Helper T cells become activated by interacting with antigen-presenting cells, such as macrophages. Antigen-presenting cells ingest a microbe, partially degrade it, and export fragments of the microbe—*i.e.,* antigens—to the cell surface, where they are presented in association with class II MHC molecules. A receptor on the surface of the helper T cell then binds to the MHC-antigen complex. But this event alone does not

Activation of T cells

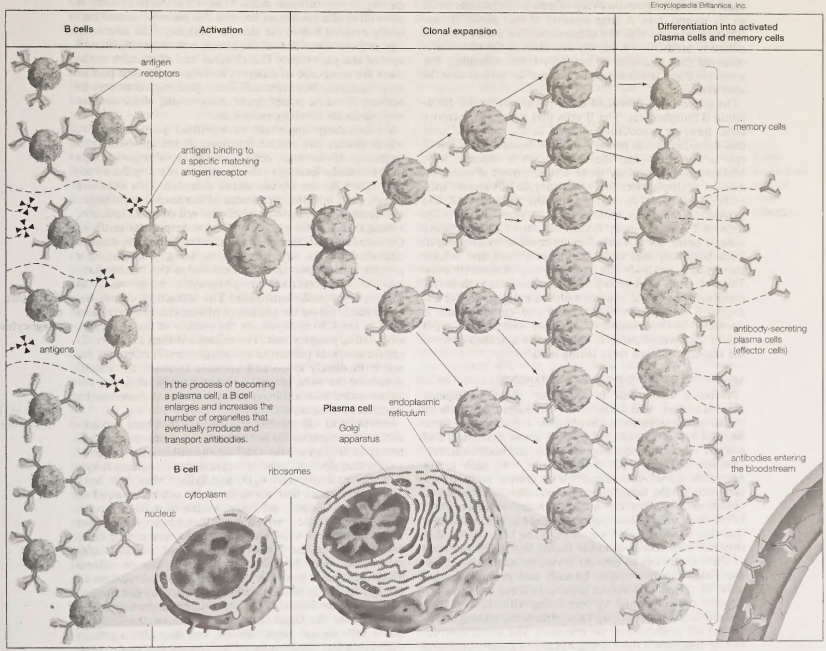
activate the helper T cell. Another signal is required, and it is provided in one of two ways: either through stimulation by a cytokine or through a costimulatory reaction between the signaling protein, B7, found on the surface of the antigen-presenting cell, and the receptor protein, CD28, on the surface of the helper T cell. If the first signal and one of the second signals are received, the helper T cell becomes activated to proliferate and to stimulate the appropriate immune cell. If only the first signal is received, the T cell may be rendered anergic—that is, unable to respond to antigen.

A discussion of helper-T-cell activation is complicated by the fact that helper T cells are not a uniform group of cells but rather can be divided into two general subpopulations—T_H1 and T_H2 cells—that have significantly different chemistry and function. These populations can be distinguished by the cytokines they secrete. T_H1 cells primarily produce the cytokines gamma interferon, tumour necrosis factor-beta, and interleukin-2 (IL-2), while T_H2 cells mainly synthesize the interleukins IL-4, IL-5, IL-6, IL-9, IL-10, and IL-13. The main role of the T_H1 cells is to stimulate cell-mediated responses (those involving cytotoxic T cells and macrophages), while T_H2 cells primarily assist in stimulating B cells to make antibodies.

Once the initial steps of activation have occurred, helper T cells synthesize other proteins, such as signaling proteins and the cell-surface receptors to which the signaling proteins bind. These signaling molecules play a critical role not only in activating the particular helper T cell but also in determining the ultimate functional role and final differentiation state of that cell. For example, the helper T cell produces and displays IL-2 receptors on its surface and also secretes IL-2 molecules, which bind to these receptors and stimulate the helper T cell to grow and divide.

The overall result of helper-T-cell activation is an increase in the number of helper T cells that recognize a specific foreign antigen, and several T-cell cytokines are produced. The cytokines have other consequences, one of which is that IL-2 allows cytotoxic or regulatory T cells that recognize the same antigen to become activated and to multiply. Cytotoxic T cells, in turn, can attack and kill other cells that express the foreign antigen in association with class I MHC molecules, which—as explained above—are present on almost all cells. So, for example, cytotoxic T cells can attack target cells that express antigens made by viruses or bacteria growing within them (see below *Cell-mediated immune mechanisms*). Regulatory T cells may be similar to cytotoxic T cells, but they are detected by their ability to suppress the action of B cells or even of helper T cells (perhaps by killing them). Regulatory T cells thus act to damp down the immune response and can sometimes predominate so as to suppress it completely.

B cells. A B cell becomes activated when its receptor recognizes an antigen and binds to it. In most cases, how- Activation of B cells



Encyclopædia Britannica, Inc.

Figure 7: *Clonal selection of a B cell.*
Activated by the binding of an antigen to a specific matching receptor on its surface, a B cell proliferates into a clone. Some clonal cells differentiate into plasma cells, which are short-lived cells that secrete antibody against the antigen. Others form memory cells, which are longer-lived and which, by proliferating rapidly, help to mount an effective defense upon a second exposure to the antigen.

ever, B-cell activation is dependent on a second factor mentioned above—stimulation by an activated helper T cell. Once a helper T cell has been activated by an antigen, it becomes capable of activating a B cell that has already encountered the same antigen. Activation is carried out through a cell-to-cell interaction that occurs between a protein called the CD40 ligand, which appears on the surface of the activated helper T cells, and the CD40 protein on the B-cell surface. The helper T cell also secretes cytokines, which can interact with the B cell and provide additional stimulation. Antigens that induce a response in this manner, which is the typical method of B-cell activation, are called T-dependent antigens.

Most antigens are T-dependent. Some, however, are able to stimulate B cells without the help of T cells. The T-independent antigens are usually large polymers with repeating, identical antigenic determinants. Such polymers often make up the outer coats and long, tail-like flagella of bacteria. Immunologists think that the enormous concentration of identical T-independent antigens creates a strong enough stimulus without requiring additional stimulation from helper T cells.

Interaction with antigens causes B cells to multiply into clones of immunoglobulin-secreting cells (see Figure 7). Then the B cells are stimulated by various cytokines to develop into the antibody-producing cells called plasma cells. Each plasma cell can secrete several thousand molecules of immunoglobulin every minute and continue to do so for several days. A large amount of that particular antibody is released into the circulation. The initial burst of antibody production gradually decreases as the stimulus is removed (for example, by recovery from infection), but some antibody continues to be present for several months afterward.

The process just described takes place among the circulating B lymphocytes. The B cells that are called memory cells, however, encounter antigen in the germinal centres—compartments in the lymphoid tissues where few T cells are present—and are activated in a different way. Memory cells, especially those with the most effective receptors, multiply extensively, but they do not secrete antibody. Instead, they remain in the tissues and the circulation for many months or even years. If, with the help of T cells, memory B cells encounter the activating antigen again, these B cells rapidly respond by dividing to form both activated cells that manufacture and release their specific antibody and another group of memory cells. The first group of memory cells behaves as though it "remembers" the initial contact with the antigen. So, for example, if the antigen is microbial and an individual is reinfected by the microbe, the memory cells trigger a rapid rise in the level of protective antibodies and thus prevent the associated illness from taking hold.

ANTIBODY-MEDIATED IMMUNE MECHANISMS

Protective attachment to antigens. Many pathogenic microorganisms and toxins can be rendered harmless by the simple attachment of antibodies. For example, some harmful bacteria, such as those that cause diphtheria and tetanus, release toxins that poison essential body cells. Antibodies, especially IgG, that combine with such toxins neutralize them. Also susceptible to simple antibody attachment are the many infectious microbes—including all viruses and some bacteria and protozoans—that live within the body cells. These pathogens bear special molecules that they use to attach themselves to the host cells so that they can penetrate and invade them. Antibodies can bind to these molecules to prevent invasion. Antibody attachment also can immobilize bacteria and protozoans that swim by means of whiplike flagella. In these instances antibodies protect simply by combining with the repeating protein units that make up these structures, although they do not kill or dispose of the microbes. The actual destruction of microbes involves phagocytosis by granulocytes and macrophages, and this is greatly facilitated by the participation of the complement system.

Activation of the complement system. "Complement" is a term used to denote a group of more than 30 proteins that act in concert to enhance the actions of other defense mechanisms of the body. Complement proteins are produced by liver cells and, in many tissues, by macrophages. Most of these proteins circulate in the blood and other body fluids in an inactive form. They become activated in sequential fashion; once the first protein in the pathway is turned on, the following complement proteins are called into action, with each protein turning on the next one in line.

The action of complement is nonspecific—*i.e.,* complement proteins are not recognized by and do not interact with antigen-binding sites. In fact, complement proteins probably evolved before antibodies. Complement functions are similar among many species, and corresponding components from one species can carry out the same functions when introduced into another species. The complement system is ingenious in providing a way for antibodies, whatever their specificity, to produce the same biological effects when they combine with antigens.

Originally immunologists thought that the complement system was initiated only by antigen-antibody complexes, but later evidence showed that other substances, such as the surface components of a microorganism alone, could trigger complement activation. Thus, there are two complement activation pathways: the first one to be discovered, the classical pathway, which is initiated by antigen-antibody complexes; and the alternative pathway, which is triggered by other means, including invading pathogens or tumour cells. (The term "alternative" is something of a misnomer because this pathway almost certainly evolved before the classical pathway. The terminology reflects the order of discovery, not the evolutionary age, of the pathways.) The classical and alternative pathways are composed of different proteins in the first part of their cascades, but eventually both pathways converge to activate the same complement components, which destroy and eliminate invading pathogens.

A particularly important complement protein is C3b, which carries out several functions. It brings about lysis (bursting) of the target cell by activating subsequent steps in the cascade, leading to the formation of a ringlike structure called the membrane attack complex. This structure inserts itself into the membrane of the invading pathogen and creates a hole through which the cell contents leak out, killing the cell. But perhaps the most important result of C3b production is that great numbers of C3b molecules are deposited on the surface of an invading pathogen in a process called opsonization. This makes the microorganism more attractive to phagocytic cells such as macrophages and neutrophils. The attraction occurs because receptors on the surface of phagocytes recognize and bind to the C3b molecule on the surface of the pathogen, stimulating phagocytosis. The microbe is then killed by digestive enzymes present in the phagocytes. If microbes are not immediately killed and are able to reach the bloodstream or the liver, spleen, or bone marrow, they can become coated with antibody and complement there and be ingested by phagocytes.

Activation of phagocytes

Activation of killer cells. Some cells that bear antigen-antibody complexes do not attract complement; their antibody molecules are far apart on the cell surface or are of a class that does not readily activate the complement system (for example, IgA, IgD, and IgE). Other cells have outer membranes that are so tough or can be repaired so quickly that the cells are impermeable to activated complement. Still others are so large that phagocytes cannot ingest them. Such cells, however, can be attacked by killer cells present in the blood and lymphoid tissues. Killer cells, which may be either cytotoxic T cells or natural killer cells, have receptors that bind to the tail portion of the IgG antibody molecule (the part that does not bind to antigen). Once bound, killer cells insert a protein called perforin into the target cell, causing it to swell and burst. Killer cells do not harm bacteria, but they play a role in destroying body cells infected by viruses and some parasites.

Other antibody-mediated mechanisms. The protection conferred by IgA antibodies, which are transported to the surface of mucous-membrane-lined passages, is somewhat different. Complement activation is not involved; there are

no complement proteins in the lining of the gut or the respiratory tract. Here the available immune defense mechanism is primarily the action of IgA combining with microbes to prevent them from entering the cells of the lining. The bound microbes are then swept out of the body. IgA also appears to direct certain types of cell-mediated killing.

IgE antibodies also invoke unique mechanisms. As stated earlier, most IgE molecules are bound to special receptors on mast cells and basophils. When antigens bind to IgE antibodies on these cells, the interaction does not cause ingestion of the antigens but rather triggers the release of pharmacologically active chemical contents of the cells' granules. The chemicals released cause a sudden increase in permeability of the local blood vessels, the adhesion and activation of platelets (blood cell fragments that trigger clotting), which release their own active agents, the contraction of smooth muscle in the gut or in the respiratory tubes, and the secretion of fluids—all of which tend to dislodge large multicellular parasites such as hookworms. Eosinophil granulocytes and IgE together are particularly effective at destroying parasites such as the flatworms that cause schistosomiasis. The eosinophils plaster themselves to the worms bound to IgE and release chemicals from their granules that break down the parasite's tough, protective skin. Therefore, IgE antibodies—although they can be a nuisance when they react with otherwise harmless antigens, as discussed in immune system disorder: Type I hypersensitivity—appear to have a special protective role against the larger parasites.

Transfer of antibodies from mother to offspring. A newborn mammal has no opportunity to develop protective antibodies on its own, unless, as happens very rarely, it was infected while in the uterus. Yet it is born into an environment similar to its mother's, that contains all the potential microbial invaders to which she is exposed. Although the fetus possesses the components of innate immunity, it has few or none of its mother's lymphocytes. The placenta generally prevents the maternal lymphocytes from crossing into the uterus, where they would recognize the fetal tissues as foreign antigens and cause a reaction similar to the rejection of an incompatible organ transplant.

What is transferred across the placenta in many species is a fair sample of the mother's antibodies. How this happens depends on the structure of the placenta, which varies among species. In humans maternal IgG antibodies—but not those of the other immunoglobulin classes—are transported across the placenta into the fetal bloodstream throughout the second two-thirds of pregnancy. In many rodents a similar transfer occurs, but primarily across the yolk sac.

In horses and cattle, which have more layers of cells in their placentas, no antibodies are transferred during fetal life, and the newborn arrives into the world with no components of specific immunity. There is, however, a second mechanism that makes up for this deficiency. The early milk (colostrum) is very rich in antibodies—mainly IgA but also some IgM and IgG—and during the first few days of life the newborn mammal can absorb these proteins intact from the digestive tract directly into the bloodstream. Drinking colostrum is therefore essential for newborn horses and cattle and required to a somewhat lesser extent by other mammals. The capacity of the digestive tract to absorb intact proteins must not last beyond one or two weeks, since once foods other than milk are ingested the proteins and other antigens in them would also be absorbed intact and could act as immunogens to which the growing animal would become allergic (see *Immune system disorders: Allergies*). IgA in milk is, however, rather resistant to digestion and can function within the gut even after intact absorption into the bloodstream has ended. Human colostrum is also rich in IgA, with the concentration highest immediately after birth.

Role of colostrum

After a newborn has received its supply of maternal antibodies, it is as fully protected as its mother. This means, of course, that if the mother has not developed immunity to a particular pathogen, the newborn will likewise be unprotected. For this reason, a physician may recommend that a prospective mother receive immunizations against tetanus and certain other disorders. (The active immunization of pregnant women against certain viral diseases, such as rubella [German measles], must be avoided, however, because the immunizing agent can cross the placenta and produce severe fetal complications.)

As important as the passively transferred maternal antibodies are, their effects are only temporary. The maternal antibodies in the blood become diluted as the animal grows; moreover, they gradually succumb to normal metabolic breakdown. Because the active development of acquired immunity is a slow and gradual process, young mammals actually become more susceptible to infection during their early stages of growth than they are immediately after birth.

Occasionally the transfer of maternal antibodies during fetal life can have harmful consequences. A well-known example of this is erythroblastosis fetalis, or hemolytic disease of the newborn, a disorder in which maternal antibodies destroy the child's red blood cells during late pregnancy and shortly after birth. The most severe form of erythroblastosis fetalis is Rh hemolytic disease, which develops when: (1) The fetus is Rh-positive; that is, its red blood cells carry an antigen known as the Rh factor. (2) The mother is Rh-negative, which is to say her red blood cells lack the Rh factor. (3) The mother's immune system has been previously activated against the Rh antigen; this usually is the result of exposure to fetal cells during the birth of an earlier Rh-positive baby or a transfusion of Rh-positive blood.

Rh hemolytic disease can be prevented by giving the mother injections of anti-Rh antibody shortly after the birth of an Rh-positive child. This antibody destroys any Rh-positive fetal cells in the maternal circulation, thereby preventing the activation of the mother's immune system should she conceive another Rh-positive fetus.

CELL-MEDIATED IMMUNE MECHANISMS

In addition to their importance in cooperating with B cells that secrete specific antibodies, T cells have important, separate roles in protecting against antigens that have escaped or bypassed antibody defenses. Immunologists have long recognized that antibodies do not necessarily protect against viral infections, because many viruses can spread directly from cell to cell and thus avoid encountering antibodies in the bloodstream. It is also known that persons who fail to make antibodies are very susceptible to bacterial infections but are not unduly liable to viral infections. Protection in these cases results from cell-mediated immunity, which destroys and disposes of body cells in which viruses or other intracellular parasites (such as the bacteria that cause tuberculosis and leprosy) are actively growing, thus depriving microorganisms of their place to grow and exposing them to antibodies.

T-cell protection against viral infection

As discussed above in *Activation of lymphocytes,* cell-mediated immunity has two mechanisms. One involves activated helper T cells, which release cytokines. In particular, the gamma interferon produced by helper T cells greatly increases the ability of macrophages to kill ingested microbes; this can tip the balance against microbes that otherwise resist killing. Gamma interferon also stimulates natural killer cells. The second mechanism of cell-mediated immunity involves cytotoxic T cells. They attach themselves by their receptors to target cells whose surface expresses appropriate antigens (notably ones made by developing viruses) and damage the infected cells enough to kill them.

Cytotoxic T cells may kill infected cells in a number of ways. The mechanism of killing used by a given cytotoxic T cell depends mainly on a number of costimulatory signals. In short, cytotoxic T cells can kill their target cells either through the use of pore-forming molecules, such as perforins and various components of cytoplasmic granules, or by triggering a series of events with the target cell that activate a cell death program, a process called apoptosis. In general, the granular cytotoxic T cells tend to kill cells directly by releasing the potent contents of their cytotoxic granules at the site of cell-to-cell contact. This renders the cell membrane of the target cell permeable, which allows the cellular contents to leak out and the cell to die (see Fig-

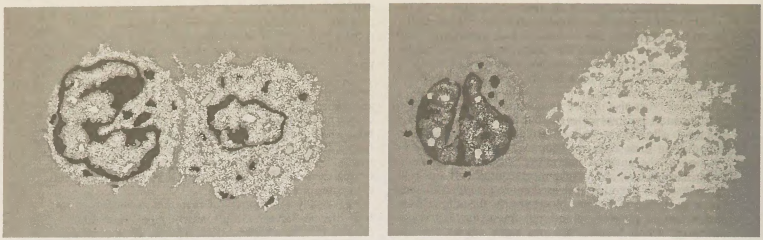


Figure 8: *The killing of an infected cell by a cytotoxic T cell.*
(Left) A scanning electron micrograph of a cytotoxic T cell, at left, attaching to a target cell infected by a virus. (Right) The target cell, on the right, is lysed, its internal contents leaking through its perforated outer membrane.
© C. Edelmann/Petit Format

ure 8). The nongranular cytotoxic T cells often kill cells by inducing apoptosis, usually through the activation of a cell-surface protein called Fas. When a protein on the surface of the cytotoxic T cell interacts with the Fas protein on the target cell, Fas is activated and sends a signal to the nucleus of the target cell, thus initiating the cell death process. The target cell essentially commits suicide, thereby destroying the virus within the cell as well.

IMMUNITY AGAINST CANCER

Cancer cells are normal body cells that have been altered in a manner that allows them to divide relentlessly, ignoring normal signals of restraint. As a result, cancer cells form clusters of cells, called tumours, that invade and colonize tissues, eventually undermining organ function and causing death. In the early 20th century the pioneering immunologist Paul Ehrlich pointed out that the enormous multiplication and differentiation of cells during prenatal life must afford many opportunities for aberrant cells to appear and grow but that immune mechanisms eliminate such cells. The idea that such a mechanism continues to function throughout life, weeding out newly arisen cancer cells, became popular in the 1950s and '60s when a number of immunologists postulated immune surveillance, the theory that T-cell-mediated immunity evolved as a specific defense against cancer cells and that T cells constantly patrol the body, searching for abnormal body cells that carry antigens on their surface which are not found on healthy body cells. Although it has its compelling aspects, the immunosurveillance theory remains just a theory, and a controversial one at that.

The role of the immune system in protecting against cancer has not been fully explained, but nevertheless there is no question that in some instances the immune system can distinguish cancer cells from normal cells. The study of tumour immunology has shown unequivocally that cancer cells do carry antigens that are not present on healthy cells. Immunologists distinguish broadly between two types of tumour antigens: tumour-specific antigens, which are found only on cancer cells and not on their normal counterparts, and tumour-associated antigens, which are found on both normal and cancer cells but which are abnormally expressed—*e.g.*, are overproduced—on cancer cells. In both cases these antigens have been shown to evoke an immune response, although not necessarily one strong enough to eliminate the tumour.

Why does a tumour continue to grow if an immune response against it is induced? Through animal experiments, a number of mechanisms have been identified that allow tumours to avoid recognition and destruction by the immune system: (1) The surfaces of cancer cells may lose antigens that are recognizable by the immune system. (2) Cancer cells may lose all class I MHC molecules from their surface, which prevents cytotoxic T cells from recognizing the cells. (3) Some cancer cells produce immunosuppressive chemicals that can inhibit T cells directly or that can activate regulatory T cells. (4) Some cancer cells shed some of their antigens, and these newly released, free-floating antigens may bind to the receptors on cytotoxic T cells, plugging them up so that the T cells cannot bind to the cancer cells and eliminate them. (5) Certain cancer cells can outmaneuver an immune response by growing so rapidly or becoming such a dense mass that immune cells cannot come in contact with most of them.

Tumour growth

Other dysfunctions of the immune system, such as immune suppression and immune deficiency, may contribute to cancer development and growth. Individuals such as transplant patients who have been treated with immunosuppressive drugs for a long period of time are more likely to develop certain types of cancer, as are patients with immunodeficiency diseases. For example, people with AIDS (acquired immunodeficiency syndrome) are more prone to developing cancers associated with viruses, such as Kaposi's sarcoma. The incidence of cancer also increases greatly in old age, when some immune responses decline. But defective immune responses may not be the major factor involved in cancer development in the elderly, since genetic mutations that are linked to cancer also accumulate with age.

Much research has been devoted to developing effective immunotherapies against cancer, but the effectiveness of this approach has been marginal. Nevertheless, researchers continue to pursue immunotherapeutic approaches. One avenue of research has focused on finding ways to immunize patients against the specific cancer growing within them. This approach targets tumour-specific antigens found on the cancer cells. Because these antigens are altered forms of normal self antigens, they are "foreign" and could be recognized by the immune system as such, but often they are not. However, investigators are working to develop vaccines that stimulate an immune response to these antigens, hoping that the reaction would be strong enough to eliminate the cancer.

PROPHYLACTIC IMMUNIZATION

Prophylactic immunization refers to the artificial establishment of specific immunity, a technique that has significantly reduced suffering and death from a variety of infectious diseases. There are two types of prophylactic immunization: passive immunization, in which protection is conferred by introducing preformed antibodies or lymphocytes from another individual whose immune system was stimulated by the appropriate antigen, and active immunization, in which protection results from the administration of a vaccine, with dead or harmless living forms of an organism or with an inactivated toxin, that stimulates the immune system to produce lymphocytes and antibodies against that organism or toxin.

Passive immunization. It is sometimes the case that an infectious organism or a poisonous substance can have such a rapid deleterious effect that the victim does not have time to develop an immune response spontaneously. At

such times passive immunization with preformed antibodies can provide life-saving assistance in combating the pathogen or poison. This situation may arise in victims of poisonous snakebites or botulism, as well as in those in whom such infections as diphtheria, tetanus, or gas gangrene have progressed to the point at which bacterial toxins have been absorbed into the bloodstream. It is also the case with bites from a rabid animal, although active immunization is begun at the same time, since the spread of the rabies infection to the central nervous system is relatively slow. Physicians use passive immunization as temporary protection for persons traveling to countries where hepatitis B is prevalent. Passive immunization provides antibodies to persons who suffer from B-cell deficiencies and are therefore unable to make antibodies for themselves (see below *Immune system disorders: Immune deficiencies*). Also, as discussed earlier, passive immunizations of anti-Rh antibody can prevent erythroblastosis fetalis.

Manufacture of antiserum

Protective immunoglobulins—primarily of the IgG class—can be prepared from the blood of humans or other species (*e.g.,* horses or rabbits) that have already developed specific immunity against the relevant antigens. These preparations are known as antiserums. (This explains the original term for passive immunization, which is serum therapy.) Human IgG is slowly broken down in the recipient's body, the concentration falling by about one-half every three weeks, so that effective amounts of antibody can be present for two or three months. Human antiserum is used whenever it is available, because IgG from other species is far more likely to provoke an immune response that will eliminate the antibody and may lead to serum sickness (see below *Immune system disorders: Type III hypersensitivity*).

Active immunization. Active immunization aims to ensure that a sufficient supply of antibodies or T and B cells that react against a potential infectious agent or toxin are present in the body before infection occurs or the toxin is encountered. Once it has been primed, the immune system either can prevent the pathogen from establishing itself or can rapidly mobilize the various protective mechanisms described above to abort the infection or toxin in its earliest stages.

The vaccines used to provide active immunization need not contain living microbes. What matters is that they include the antigens important in evoking a protective response and that those antigens be administered in a harmless form sufficient in amount and persistence to produce an immune response similar to the natural infection. Bacterial toxins, such as those that cause tetanus or diphtheria, can be rendered harmless by treatment with formaldehyde without affecting their ability to act as immunogens. These modified toxins, or toxoids, usually are adsorbed onto an inorganic gel before being administered, an approach that increases the likelihood that the toxoid will be retained in a macrophage. Toxoids elicit effective, long-lasting immunity against bacterial toxins. When immunization against several antigenic determinants is desired or the important antigenic component is not known, it may be prudent to use the entire microbe, which has been killed in a manner that does not alter it significantly. Such so-called "killed" vaccines are used to immunize against typhoid, pertussis (whooping cough), plague, and influenza, for example. In other cases, researchers have developed attenuated (*i.e.,* weakened) strains of bacteria or viruses. Attenuated vaccines cause an infection but do not produce the full array of signs and symptoms of the disease, because the infectious agent multiplies to only a limited extent in the body and never reverts to the virulent form. The use of such live microbes provides the most effective prophylaxis of all, since they truly imitate a mild form of the natural infection. Such are the vaccines for yellow fever, poliomyelitis (oral vaccine), measles, rubella, and tuberculosis. Although sufficiently attenuated as far as healthy persons are concerned, live vaccines may cause the full disease in persons who have an immune deficiency.

Types of vaccines

Most vaccines are administered by injection, but a few are given orally. Ultimately mucosal vaccines (those administered to mucosal surfaces such as those lining the gut, nasal passages, or the urogenital tract) may be the most effective vaccines available because of their unique ability to stimulate IgA responses and because of their ease of administration. Recombinant DNA technology has allowed researchers to use modified bacteria and viruses that are not harmful to humans to immunize individuals against an antigen from a pathogenic microorganism. This approach involves introducing into the DNA of the harmless microorganism a gene from a pathogenic organism that encodes an antigen capable of eliciting a protective immune response but not the full-blown disease. Once inoculated into the host, the microorganism generates the protective antigen of the pathogen and immunizes the host. An effective oral vaccine against cholera was developed based on this approach.

Sometimes different strains of a microorganism, each characterized by a different antigenic determinant, give rise to the same disease. In such cases neither natural infection nor prophylactic immunization with any one strain protects against infection by the others. For example, a variety of virus strains cause the common cold, but it is impractical to immunize against each. On the other hand, although there are more than 60 different strains of pneumococci that can cause bacterial pneumonia, some are much more common than others. Consequently a vaccine containing antigens from up to 14 of the most common strains is useful in protecting persons at special risk.

Active immunization is often the most effective and least costly method of protecting against an infectious disease. Vaccination campaigns against many diseases, such as diphtheria, polio, and measles, have been tremendously successful (see Figure 9). In cases in which 95 percent or

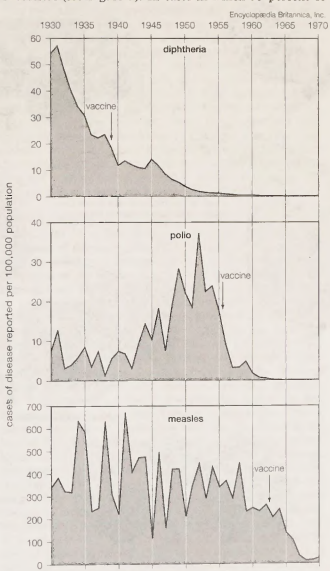


Figure 9: In the United States, mass vaccination programs carried out against diphtheria, polio, and measles have almost eradicated these diseases from the population. The graphs indicate the years the vaccines were introduced. Data source: U.S. Bureau of the Census, *Historical Statistics of the United States: Colonial Times to 1970* (CD-ROM ed., 1997).

more of the population at risk is protected and humans are the only reservoir of infection, active immunization can lead to the worldwide eradication of the infectious agent, as has been achieved in the case of smallpox.

PRODUCTION OF MONOCLONAL ANTIBODIES

It was pointed out above that upon activation by an antigen, a circulating B cell multiplies to form a clone of plasma cells, each secreting the identical immunoglobulin. These immunoglobulins—derived from the descendants of a single B cell—constitute a monoclonal antibody. The complete immune response to a natural infection or an active immunization, however, is polyclonal. In other words, it involves many B-cell clones, each of which recognizes a different antigenic determinant and secretes a different immunoglobulin. Thus the blood serum of an immunized person or animal normally contains a mixture of antibodies. There is, however, a condition in which the blood serum may contain an astonishingly high concentration of a single immunoglobulin. This results from multiple myeloma, a type of cancer in which a single B cell proliferates to form a tumorous clone of antibody-secreting cells that can multiply indefinitely, like all cancer cells. Myelomas have been propagated to produce large quantities of monoclonal antibodies, and the use of myeloma cells to produce specific, desired antibodies has become one of the most important facets of biotechnology.

Use of myeloma cells

The artificial production of monoclonal antibodies from cultured myeloma cells was pioneered in 1975 by immunologists Georges Köhler and César Milstein (who received Nobel Prizes in 1984 for their work). The basic process, as shown in Figure 10, begins with the selection of myeloma cells that grow well but have lost their ability to secrete immunoglobulin and do not produce a crucial enzyme called HGPRT (hypoxanthineguanine phosphoribosyltransferase). These cells are induced to fuse with a set of plasma cells, taken from the spleen of an inoculated mouse, that is known to secrete a particular antibody. The result is a "hybridoma," or hybrid myeloma, which retains the capacity of its myeloma component to multiply indefinitely but also makes the chosen, identifiable antibody of its plasma-cell component. The fused mixture is placed into a medium known as HAT (hyposanthineguanine phosphoribosyltransferase). The HAT medium kills unfused myeloma cells, which do not produce the HGPRT enzyme, and the naturally short-lived unfused plasma cells also die off. Only the hybridomas live in the HAT medium, and these are grown in bulk to produce the specified antibodies.

Thanks to hybridomas, researchers can obtain monoclonal antibodies that recognize individual antigenic sites on almost any molecule, ranging from drugs and hormones to microbial antigens and cell receptors. The exquisite specificity of monoclonal antibodies and their availability in quantity have made it possible to devise sensitive assays for an enormous range of biologically important substances and to distinguish cells from one another by identifying previously unknown marker molecules on their surfaces. Moreover, if short-lived radioactive atoms are added to these antibodies and they are then administered in tiny quantities to a patient, they become attached

Encyclopædia Britannica, Inc.

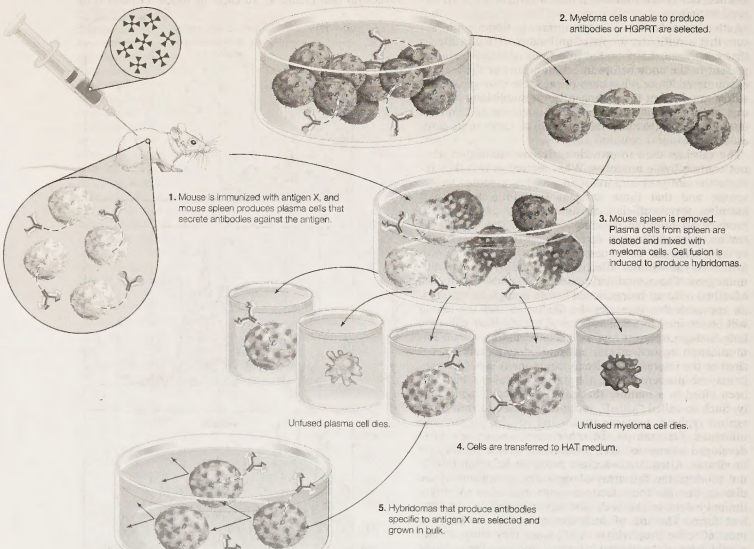


Figure 10: *Artificial production of monoclonal antibodies.*
The technique involves fusing certain myeloma cells (cancerous B cells), which can multiply indefinitely but cannot produce antibodies, with plasma cells (noncancerous B cells), which are short-lived but produce a desired antibody. The resulting hybrid cells, called hybridomas, grow at the rate of myeloma cells but also produce large amounts of the desired antibody. In this way researchers obtain large quantities of antibody molecules that all react against the same antigen.

exclusively to the cancer tissue. By means of instruments that detect the radioactivity, physicians can locate the cancerous sites without surgical intervention. Monoclonal antibodies have also been used experimentally to deliver cytotoxic drugs or radiation to cancer cells.

Although the preparation of monoclonal antibodies from rat or mouse cells has become routine practice, the construction of human hybridomas has not proved easy. This is partly because most human myeloma cells do not grow well in culture, and those that do so have not produced stable hybridomas. If, however, human B cells isolated from blood are infected by the Epstein-Barr virus (the agent that causes infectious mononucleosis), they can be propagated in culture and continue to secrete immunoglobulin. Very few of them are likely to be making an antibody with a desired specificity, even in a subject who has been immunized; but in some instances immunologists have succeeded in identifying and selecting cells that secrete the wanted immunoglobulin. These can be grown in culture as single clones that secrete a monoclonal antibody. Researchers have used this process to obtain human monoclonal antibodies against the Rh antigen.

Recombinant DNA techniques

A simpler method of constructing human monoclonal antibodies can be accomplished using recombinant DNA techniques. Once a mouse monoclonal antibody has been constructed using the traditional methods just described, DNA encoding the antigen-binding portion of the antibody molecule can be isolated and fused to human DNA that encodes an antibody. Then the hybrid DNA is inserted into a bacterium, which produces half-mouse–half-human monoclonal antibodies. The antibodies made by this method are less likely to induce an anti-antibody response when given to humans. Further fine-tuning can be done to change all parts of the antibody that are not directly involved in binding to the specific antigen. This technique has been used to produce a large number of different monoclonal antibodies for use in therapy.

Evolution of the immune system

Virtually all organisms have at least one form of defense that helps repel disease-causing organisms. Advanced vertebrate animals, a group that includes humans, defend themselves against such microorganisms by means of a complex group of defense responses collectively called the immune system. This protective system evolved from simpler defense mechanisms, but the evolutionary twists and turns that led to its development are not entirely clear. To unravel the path that the vertebrate immune system followed in its evolution, investigators have studied the defense responses of various living organisms. They also have examined the genes of immune system proteins for clues to the genetic origins of immunity.

DEVELOPMENT OF IMMUNITY IN MAJOR ANIMAL GROUPS

Because the immune system is composed of cells and tissues that do not lend themselves to fossilization, it is impossible to trace the evolution of immunity from the paleontological record. But, because all animals exhibit some general ability to recognize self and to repel foreign substances, it is possible to study the immune capacity of living animals and, based on the relative positions of these animals in the evolutionary tree, to extrapolate a reasonable evolutionary history of the immune system.

Primitive immune systems

Immune capacity among invertebrates. From the lowliest protozoans to the higher marine tunicates, invertebrates have means of distinguishing self components from nonself components. Sponges from one colony will reject tissue grafts from a different colony but will accept grafts from their own. When tissue grafts are made in animals higher up the evolutionary tree—between individual annelid worms or starfish, for example—the foreign tissue is commonly invaded by phagocytic cells (cells that engulf and destroy foreign material) and cells resembling lymphocytes (white blood cells of the immune system), and it is destroyed. Yet tissues grafted from one part of the body to another on the same individual adhere and heal readily and remain healthy. So it seems that something akin to cellular immunity is present at this level of evolution.

Insects engulf and eliminate foreign invaders through the process of phagocytosis ("cellular eating"). They have factors present in their circulatory fluids that can bind to foreign cells and cause clumping, or agglutination, of a number of these cells, an event that facilitates phagocytosis. Insects also seem to acquire immunity to infectious agents.

Immune capacity among vertebrates. The most sophisticated immune systems are those of the vertebrates. Recognizable lymphocytes and immunoglobulins (Ig; also called antibodies) appear only in these organisms. The most primitive living vertebrates—the jawless fishes (hagfish and lampreys)—do not have lymphoid tissues corresponding to a spleen or a thymus, and their immune responses, although demonstrable, are very weak and sluggish. Farther up the evolutionary tree, at the level of the cartilaginous fishes (sharks and rays) and the bony fishes, a thymus and a spleen are present, as are immunoglobulins, although only those immunoglobulins of the IgM class are detectable. Fish lack specialized lymph nodes, but they do have clusters of lymphocytes in the gut that may serve an analogous purpose.

It is not until the level of the terrestrial vertebrates—amphibians, reptiles, birds, and mammals—that a complete immune system with thymus, spleen, bone marrow, and lymph nodes is present and IgM and IgG antibodies are made. Antibodies of the IgA class are found only in birds and mammals, and IgE antibodies are confined to mammals. So it appears that the most primitive devices for producing specific, acquired immunity gradually diversified to meet the new environmental hazards as animals moved out of the sea onto the land.

The evolution of the complement system (a group of proteins involved in immune responses) may have occurred faster than that of the immunoglobulin system. The jawless fishes have complement components corresponding only to the later-acting (*i.e.*, cytolytic, or cell-killing) aspects of complement function, but all higher vertebrates have components similar to the complete complement system of mammals. The fact that the complement system has been so well conserved during evolution implies not only that it has been of great biological value but also that complement and immunoglobulins have interacted throughout the evolution of the immune system in higher vertebrates. (For more information on the complement system, see above *Antibody-mediated immune mechanisms*).

GENETIC ORIGINS OF THE IMMUNE SYSTEM

Researchers have found many similarities between the structures of proteins involved in antigen recognition and those in cell-to-cell recognition in the immune system. (Antigens are the foreign proteins that antibodies recognize and bind to.) These proteins include the antigen receptors of lymphocytes, the major histocompatibility complex (MHC) proteins, the coreceptors involved in cell-to-cell recognition in immune reactions (such as the receptors named CD4, CD8, and CD28), and the F_c receptor that binds to the stem of the Y-shaped immunoglobulin molecule. A number of proteins not involved in the immune system also share structural features with these proteins. The main feature similarity is a structure called the immunoglobulin domain. Each protein is composed of one or more Ig domains of nearly identical size. The domains are formed into a loop by bonds between sulfur atoms on the amino acids at the ends. Although each domain is different and serves a different function in the molecule as a whole, the number and order of the amino acids forming each domain are far more similar than would be expected if each had arisen independently in the course of evolution. Equally remarkable is the fact that nerve cells, thymus cells, and T lymphocytes in mice and rats carry a surface protein termed Thy-1 (thymus-1 antigen), the function of which is unknown, that also has this same basic structure and a similar arrangement of amino acids. The similarities suggest that the genes for all these molecules originated from some primitive gene involved in the recognition of one cell by another, which is required for orderly development of a complex, multicellular organism, and that during evolution they had acquired different functions. Researchers

named this group of genes and their protein products the immunoglobulin superfamily. The processes whereby one ancestral gene could have given rise to such a family of genes include gene duplication, crossing over, and mutation, all of which are discussed in detail in GENETICS AND HEREDITY, THE PRINCIPLES OF.

Not surprisingly, molecules that have a similar function in different species (*e.g.*, immunoglobulins or MHC components) show an even closer resemblance. By analyzing the number of differences in the amino acids and their position in the polypeptide chains that make up IgM and IgG molecules in, for example, humans, mice, and rabbits and by making reasonable assumptions about mutation rates, scientists can estimate roughly how many generations would have had to elapse—and hence how much time—for the present immunoglobulins of these species to have evolved from a common ancestral IgM-like molecule. Such calculations suggest that divergence from the ancestral immunoglobulin took place some 200 million years ago. That was about the same time amphibians are thought to have diverged from the main vertebrate line. So one may conclude that a functional immune system arose even earlier and has continued to provide a defense against foreign agents ever since.

IMMUNE SYSTEM DISORDERS

Disorders of the immune system include immune deficiency diseases, such as AIDS, that arise because of a diminution of some aspect of the immune response. Other types of immune disorders, such as allergies and autoimmune disorders, are caused when the body develops an inappropriate response to a substance—either to a normally harmless foreign substance found in the environment, in the case of allergies, or to a component of the body, in the case of autoimmune diseases. Finally, lymphocytes (white blood cells of the immune system) can become cancerous and give rise to tumours called leukemias, lymphomas, and myelomas. All four types of disorders—immune deficiencies, allergies, autoimmune diseases, and cancers—are discussed in this section.

Immune deficiencies

Immune deficiency disorders result from defects that occur in immune mechanisms. The defects arise in the components of the immune system, such as the white blood cells involved in immune responses (T and B lymphocytes and scavenger cells) and the complement proteins, for a number of reasons. Some deficiencies are hereditary and result from genetic mutations that are passed from parent to child. Others are caused by developmental defects that occur in the womb. In some cases immune deficiencies result from damage inflicted by infectious agents. In others drugs used to treat certain conditions, or even the diseases themselves, can depress the immune system. Poor nutrition also can undermine the immune system.

Genetic immune deficiencies

Immune deficiencies resulting from hereditary and congenital defects are rare, but they can affect all major aspects of the immune system. Luckily many of those conditions can be treated. In the rare hereditary disorder called X-linked infantile agammaglobulinemia, which affects only males, B lymphocytes are unable to secrete all classes of immunoglobulins. (An immunoglobulin is a type of protein, also called an antibody, that is produced by B cells in response to the presence of a foreign substance called an antigen.) The disease can be treated by periodic injections of large amounts of immunoglobulin G (IgG). The congenital, but not hereditary, T-cell deficiency disease called DiGeorge's syndrome arises from a developmental defect occurring in the fetus that results in the defective development of the thymus. Consequently the infant has either no mature T cells or very few. In the most severe cases—*i.e.*, when no thymus has developed—treatment of DiGeorge's syndrome consists of transplantation of a fetal thymus into the infant. The group of disorders called severe combined immunodeficiency diseases result from a failure of precursor cells to differentiate into T or B cells. Bone marrow transplantation can successfully treat some of those diseases. The immune disorder called chronic granulomatous disease results from an inherited defect that prevents phagocytic cells from producing enzymes needed to break down ingested pathogens. Treatments include administration of a wide spectrum of antibiotics.

Damage to lymphocytes that is inflicted by viruses is common but usually transient. During infectious mononucleosis, for example, the Epstein-Barr virus infects B cells, causing them to express viral antigens. T cells that react against these antigens then attack the B cells, and a temporary deficiency in the production of new antibodies lasts until the overt viral infection has been overcome. Because antibodies already present in the blood are slowly broken down, failure to make new ones is important only if the infection persists for a long time, as it does occasionally. A much more serious viral infection is that caused by the human immunodeficiency virus (HIV), which is responsible for the fatal immune deficiency disease AIDS. HIV selectively infects helper T cells and prevents them from producing cytokines and from functioning in cell-mediated immunity (see Figure 11). Persons with AIDS may be unable to overcome infections by a variety of microbes that are easily disposed of by persons uninfected with HIV. Severe infections by certain parasites, such as trypanosomes, also cause immune deficiency, as do some forms of cancer, but it is uncertain how that comes about. For example, Hodgkin's disease, which attacks the lymphatic system, makes the patient more susceptible to infection.

In countries with advanced medical services, immune deficiency often results from the use of powerful drugs to treat cancers. The drugs work by inhibiting the multiplica-

© NIBSC, Science Photo Library/Photo Researchers, Inc.

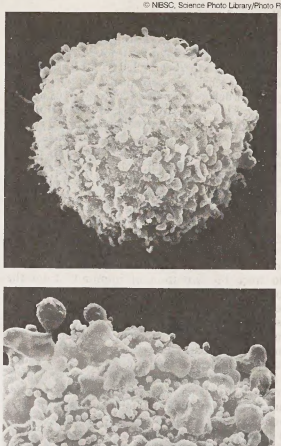


Figure 11: (Top) Scanning electron micrograph of a T lymphocyte infected with the human immunodeficiency virus (HIV). (Bottom) Close-up view of an infected T cell, showing HIV particles budding from its surface.

tion of rapidly dividing cells. Although the drugs act selectively on cancer cells, they also can interfere with the generation and multiplication of cells involved in immune responses. Prolonged or intensive treatment with such drugs impairs immune responses to some degree. Although the immune impairment is reversible, the physician must seek a balance between intentional damage to the cancer cells and unintentional damage to the immune system.

Deficiencies caused by drug therapy

Medically induced suppression of the immune system also occurs when powerful drugs, which are designed to interfere with the development of T and B cells, are used to prevent the rejection of organ or bone marrow transplants or to damp down serious autoimmune responses. Although use of such drugs has greatly improved the success of transplants, it also leaves patients highly susceptible to microbial infections. Fortunately, most of those infections can be treated with antibiotics, but the immunosuppressive drugs have to be used with great care and for as short a period as possible.

In countries where the diet, especially that of growing children, is grossly deficient in protein, severe malnutrition ranks as an important cause of immune deficiency. Antibody responses and cell-mediated immunity are seriously impaired, probably because of atrophy of the thymus and the consequent deficiency of helper T cells.

Allergies

The immune system recognizes and responds to almost any foreign molecule; it cannot discern between molecules that are characteristic of potentially infective agents and those that are not. In other words, an immune response can be induced by materials that have nothing to do with infection. The mechanisms brought into play, though beneficial for eliminating microbes, are not necessarily beneficial when otherwise innocuous substances are targeted. Furthermore, even initially protective mechanisms can cause secondary disorders when they operate on too great a scale or for a longer period than necessary, thereby damaging tissues remote from the infection. The terms "allergy" and "hypersensitivity" are commonly used to describe inappropriate immune responses that occur when an individual becomes sensitized to harmless substances. Allergic reactions do not as a rule cause symptoms to arise on the first exposure to an antigen. At the initial exposure reactive lymphocytes are generated that go into action only when the individual is reexposed to the antigen.

The manifestations of a particular allergic reaction depend on which of the immune mechanisms predominates in the response. Based on this criterion, immunologists use the Gell-Coombs classification system to recognize four types of hypersensitivity reactions. Types I, II, and III involve antibody-mediated mechanisms and are of rapid onset. The type IV reaction stems from cell-mediated mechanisms and has a delayed onset. It should be noted that the categorization, though useful, is an oversimplification and that many diseases involve a combination of hypersensitivity reactions.

Types of hypersensitivity reactions

TYPE I HYPERSENSITIVITY

Type I, also known as atopic or anaphylactic, hypersensitivity, involves IgE antibody, mast cells, and basophils.

Sensitization, activation, and effector phases. Type I hypersensitivity can be divided into three phases. The first is called the sensitization phase and occurs when the individual is first exposed to antigen. Exposure stimulates the production of IgE antibodies, which bind to mast cells and circulating basophils. The mast cells are found in tissues, often near blood vessels. The second phase is the activation phase, and it occurs when the individual is reexposed to the antigen. Reintroduction of the antigen causes IgE molecules to become cross-linked, which triggers the mast cells and basophils to release the contents of their granules into the surrounding fluids, initiating the third phase, called the effector phase, of the type I reaction. The effector phase includes all the body's complex reactions to the potent chemicals from the granules. The chemicals include histamine, which causes small blood vessels to dilate and smooth muscle in the bronchial tubes of the lungs to constrict; heparin, which prevents blood coagulation; enzymes that break down proteins; signaling agents that attract eosinophils and neutrophils; and a chemical that stimulates platelets to adhere to blood vessel walls and to release serotonin, which constricts arteries. In addition the stimulated mast cells make chemicals (prostaglandins and leukotrienes) that have potent local effects; they cause capillary blood vessels to leak, smooth muscles to contract, granulocytes to move more actively, and platelets to become sticky.

Type I allergic reactions. The overall result of the type I reaction is an acute inflammation marked by local seepage of fluid from and dilation of the blood vessels, followed by ingress of granulocytes into the tissues. This inflammatory reaction can be a useful local protective mechanism. If, however, it is triggered by an otherwise innocuous antigen entering the eyes and nose, it results in swelling and redness of the linings of the eyelids and nasal passages, secretion of tears and mucus, and sneezing—the typical symptoms of hay fever. If the antigen penetrates the lungs, not only do the linings of the bronchial tubes become swollen and secrete mucus, but the muscle in their walls contracts and the tubes are narrowed, making breathing particularly difficult. These are the symptoms of acute asthma. If the antigen is injected beneath the skin—for example, by the sting of an insect or in the course of some medical procedure—the local reaction may be extensive. Called a wheal-and-flare reaction, it includes swelling, produced by the release of serum into the tissues (wheal), and redness of the skin, resulting from the dilation of blood vessels (flare). If the injected antigen enters the bloodstream and interacts with basophils in the blood as well as with mast cells deep within the tissues, the release of active agents can cause hives, characterized by severe itching. If the antigen enters through the gut, the consequences can include painful intestinal spasms and vomiting. Local reaction with mast cells increases the permeability of the mucosa of the gut, and in many cases the antigen enters the bloodstream and also produces hives. Regardless of whether the allergen is injected or ingested, if it ends up in the bloodstream, it can induce anaphylaxis, a syndrome that in its most severe form is characterized by a profound and prolonged drop in blood pressure accompanied by difficulty in breathing. Death can occur within minutes unless an injection of epinephrine is administered immediately. This type of severe allergic reaction can occur in response to foods, drugs such as penicillin, and insect venom.

Anaphylaxis

Another feature of type I hypersensitivity reactions is that, once the immediate local reaction to the allergen has taken its course, there may occur an influx of more granulocytes, lymphocytes, and macrophages at the site. If the allergen is still present, a more prolonged form of the same reaction—the so-called late-phase reaction, which lasts a day or two rather than minutes—may supervene. This is a feature of asthmatic attacks in some subjects, in whom repeated episodes also lead to increased sensitivity of the air passages to the constrictive action of histamine. If such persons can escape exposure to the allergen for several weeks, subsequent exposure causes much less severe attacks. A prolonged IgE-induced reaction also causes atopic dermatitis, a skin condition characterized by persistent itching and scaly red patches. These often develop at sites where the skin is bent, such as the elbows and knees. The persistence is due to the influx of mast cells stimulated by the continued presence of the allergen, which is often a harmless substance such as animal hair or dander.

Typical type I allergens. Most people are not unduly susceptible to hay fever or asthma. Those who are—about 10 percent of the population—are sometimes described as atopic (from the term "atopy," meaning "uncommon"). Atopic individuals have an increased tendency to make IgE antibodies. This tendency runs in families, though there is no single gene responsible as there is in some hereditary diseases such as hemophilia. Although many innocuous antigens can stimulate a small amount of IgE antibody in the atopic individual, some antigens are much more likely to do so than others, especially if they are repeatedly absorbed in very small amounts through mucosal surfaces.

Such antigens are often termed allergens. These substances are usually polypeptides that have carbohydrate groups attached to them. They are resistant to drying, but no special characteristic is known that clearly distinguishes allergens from other antigens. Allergens are present in many types of pollen (which accounts for the seasonal incidence of hay fever), in fungal spores, in animal dander and feathers, in plant seeds (especially when finely ground) and berries, and in what is called house dust. The main allergen in house dust has been identified as the excreta of mites that live on skin scales (see Figure 12); other mites (those that live in flour, for example) also excrete potent allergens. This list is far from exhaustive. In addition to those previously named, sensitivities to chocolate, egg whites, oranges, or cow's milk are not uncommon.

© David Scharf/Peter Arnold, Inc.

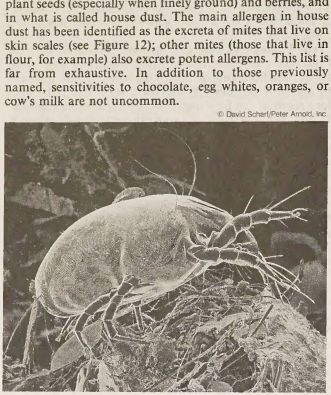


Figure 12: Scanning electron micrograph of a dust mite (*Dermato phoides*) on dust.

The amount of allergen needed to trigger an acute type I hypersensitivity reaction in a sensitive person is very small: less than one milligram can produce fatal anaphylaxis if it enters the bloodstream. Medical personnel should inquire about any history of hypersensitivity before administering drugs by injection, and if necessary they should inject a test dose into (rather than through) the skin to ensure that hypersensitivity is absent. In any case, a suitable remedy should be at hand.

Treatment of type I allergic responses. Several drugs are available that mitigate the effects of IgE-induced allergic reactions. Some, such as the anti-inflammatory cromolyn, prevent mast-cell granules from being discharged if administered before reexposure to antigen. For treatment of asthma and severe hay fever, such drugs are best administered by inhalation. The effects of histamine can be blocked by antihistamine agents that compete with histamine for binding sites on the target cells. Antihistamines are used to control mild hay fever and such skin manifestations as hives, but they tend to make people sleepy. Epinephrine counteracts, rather than blocks as antihistamines do, the effects of histamine and it is most effective in treating anaphylaxis. Corticosteroid drugs can help control persistent asthma or dermatitis, probably by diminishing the inflammatory influx of granulocytes, but long-continued administration can produce dangerous side effects and should be avoided.

Blocking the effect of histamine

Sensitivity to allergens often diminishes with time. One explanation is that increasing amounts of IgG antibodies are produced, which preferentially combine with the allergen and so prevent it from reacting with the cell-bound IgE. This is the rationale for desensitization treatment, in which small amounts of the allergen are injected beneath the skin in gradually increasing quantities over a period of several weeks, so as to stimulate IgG antibodies. The method is often successful in diminishing hypersensitivity to a tolerable level or even abolishing it. However, increased IgG production may not be the complete explanation. The capacity to make IgE antibodies depends on the cooperation of helper T cells, and they in turn are regulated by regulatory T cells. There is evidence suggesting that atopic individuals are deficient in regulatory T cells whose function is specifically to depress the B cells that produce IgE and that desensitization treatment may overcome this deficiency.

TYPE II HYPERSENSITIVITY

Allergic reactions of this type, also known as cytotoxic reactions, occur when cells within the body are destroyed by antibodies, with or without activation of the entire complement system. When antibody binds to an antigen on the surface of a target cell, it can cause damage through a number of mechanisms. When IgM or IgG molecules are involved, they activate the complete complement system, which leads to the formation of a membrane attack complex that destroys the cell (see above *Antibody-mediated immune mechanisms*). Another mechanism involves IgG molecules, which coat the target cell and attract macrophages and neutrophils to destroy it. Unlike type I reactions, in which antigens interact with cell-bound IgE immunoglobulins, type II reactions involve the interaction of circulating immunoglobulins with cell-bound antigens.

Type II reactions only rarely result from the introduction of innocuous antigens. More commonly, they develop because antibodies have formed against body cells that have been infected by microbes (and thus present microbial antigenic determinants) or because antibodies have been produced that attack the body's own cells. This latter process underlies a number of autoimmune diseases, including autoimmune hemolytic anemia, myasthenia gravis, and Goodpasture's syndrome.

Reactions against blood transfusions

Type II reactions also occur after an incompatible blood transfusion, when red blood cells are transfused into a person who has antibodies against proteins on the surface of these foreign cells (either naturally or as a result of previous transfusions). Such transfusions are largely avoidable, but when they do occur the effects vary according to the class of antibodies involved. If these activate the complete complement system, the red cells are rapidly hemolyzed (made to burst), and the hemoglobin in them is released into the bloodstream. In small amounts it is mopped up by a special protein called hemopexin, but in large amounts it is excreted through the kidneys and can damage the kidney tubules. If activation of complement only goes part of the way (to the C3 stage), the red cells are taken up and destroyed by granulocytes and macrophages, mainly in the liver and spleen. The heme pigment from the hemoglobin is converted to the pigment bilirubin, which accumulates in the blood and makes the person appear jaundiced.

Not all type II reactions cause cell death. Instead the antibody may cause physiological changes underlying disease. This occurs when the antigen to which the antibody binds is a cell-surface receptor, which normally interacts with a chemical messenger, such as a hormone. If the antibody binds to the receptor, it prevents the hormone from binding and carrying out its normal cellular function (see below *Autoimmune diseases of the thyroid gland*).

TYPE III HYPERSENSITIVITY

Type III, or immune-complex, reactions are characterized by tissue damage caused by the activation of complement in response to antigen-antibody (immune) complexes that are deposited in tissues. The classes of antibody involved are the same ones that participate in type II reactions—IgG and IgM—but the mechanism by which tissue damage is brought about is different. The antigen to which the antibody binds is not attached to a cell. Once the antigen-antibody complexes form, they are deposited in various tissues of the body, especially the blood vessels, kidneys, lungs, skin, and joints. Deposition of the immune complexes causes an inflammatory response, which leads to the release of tissue-damaging substances, such as enzymes that destroy tissues locally, and interleukin-1, which, among its other effects, induces fever.

Immune complexes underlie many autoimmune diseases, such as systemic lupus erythematosus (an inflammatory disorder of connective tissue), most types of glomerulonephritis (inflammation of the capillaries of the kidney), and rheumatoid arthritis.

Type III hypersensitivity reactions can be provoked by in-

halation of antigens into the lungs. A number of conditions are attributed to this type of antigen exposure, including farmer's lung, caused by fungal spores from moldy hay; pigeon fancier's lung, resulting from proteins from powdery pigeon dung; and humidifier fever, caused by normally harmless protozoans that can grow in air-conditioning units and become dispersed in fine droplets in climate-controlled offices. In each case, the farmer, pigeon fancier, or office worker will be sensitized to the antigen—*i.e.*, will have IgG antibodies to the agent circulating in the blood. Inhalation of the antigen will stimulate the reaction and cause chest tightness, fever, and malaise, symptoms that usually pass in a day or two but recur when the individual is reexposed to the antigen. Permanent damage is rare unless individuals are exposed repeatedly. Some occupational diseases of workers who handle cotton, sugarcane, or coffee waste in warm countries have a similar cause, with the sensitizing antigen usually coming from fungi that grow on the waste rather than the waste itself. The effective treatment is, of course, to prevent further exposure.

Serum sickness

The type of allergy described in the preceding paragraph was first recognized as serum sickness, a condition that often occurred after animal antiserum had been injected into a patient to destroy diphtheria or tetanus toxins. While still circulating in the blood, the foreign proteins in the antiserum induced antibodies, and some or all of the symptoms described above developed in many subjects. Serum sickness is now rare, but similar symptoms can develop in people sensitive to penicillin or certain other drugs, such as sulfonamides. In such cases the drug combines with the subject's blood proteins, forming a new antigenic determinant to which antibodies react.

The consequences of antigen-and-antibody interaction within the bloodstream vary according to whether the complexes formed are large, in which case they are usually trapped and removed by macrophages in the liver, spleen, and bone marrow, or small, in which case they remain in the circulation. Large complexes occur when more than enough antibody is present to bind to all the antigen molecules, so that these form aggregates of many antigen molecules cross-linked together by the multiple binding sites of IgG and IgM antibodies. When the ratio of antibody to antigen is enough to form only small complexes, which can nevertheless activate complement, the complexes tend to settle in the narrow capillary vessels of the synovial tissue (the lining of joint cavities), the kidney, the skin, or, less commonly, the brain or the mesentery of the gut. The activation of complement—which leads to increased permeability of the blood vessels, release of histamine, stickiness of platelets, and attraction of granulocytes and macrophages—becomes more important when the antigen-antibody complexes are deposited in blood vessels than when they are deposited in the tissues outside the capillaries. The symptoms, depending on where the damage occurs, are swollen, painful joints, a raised skin rash, nephritis (kidney damage, causing blood proteins and even red blood cells to leak into the urine), diminished blood flow to the brain, or gut spasms.

The formation of troublesome antigen-antibody complexes in the blood can also result from subacute bacterial endocarditis, a chronic infection of damaged heart valves. The infectious agent is often *Streptococcus viridans*, normally a harmless inhabitant of the mouth. The bacteria in the heart become covered with a layer of fibrin, which protects them from destruction by granulocytes, while they continue to release antigens into the circulation. These can combine with preformed antibodies to form immune complexes that can cause symptoms resembling those of serum sickness.

TYPE IV HYPERSENSITIVITY

Type IV hypersensitivity is a cell-mediated immune reaction. In other words, it does not involve the participation of antibodies but is due primarily to the interaction of T cells with antigens. Reactions of this kind depend on the presence in the circulation of a sufficient number of T cells able to recognize the antigen. The specific T cells must migrate to the site where the antigen is present. Since this process takes more time than reactions involving antibodies, type IV reactions first were distinguished by their delayed onset and are still frequently referred to as delayed hypersensitivity reactions. Type IV reactions not only develop slowly—reactions appear about 18 to 24 hours after introduction of antigen to the system—but, depending on whether the antigen persists or is removed, they can be prolonged or relatively transient.

The T cells involved in type IV reactions are memory cells derived from prior stimulation by the same antigen. These cells persist for many months or years, so that persons who have become hypersensitive to an antigen tend to remain so. When T cells are restimulated by this antigen presented on the surface of the macrophages (or on other cells that can express class II MHC molecules), the T cells secrete cytokines that recruit and activate lymphocytes and phagocytic cells, which carry out the cell-mediated immune response. Two common examples of delayed hypersensitivity that illustrate the various consequences of type IV reactions are tuberculin-type and contact hypersensitivity.

Tuberculin-type hypersensitivity. The tuberculin test is based on a delayed hypersensitivity reaction. The test is used to determine whether an individual has been infected with the causative agent of tuberculosis, *Mycobacterium tuberculosis*. (A previously infected individual would harbour reactive T cells in the blood.) In this test, small amounts of protein extracted from the mycobacterium are injected into the skin. If reactive T cells are present—*i.e.*, the test is positive—redness and swelling appear at the injection site the next day, increase through the following day, and then gradually fade away. If a tissue sample from the site of the positive reaction is examined, it will show infiltration by lymphocytes and monocytes, increased fluid between the fibrous structures of the skin, and some cell death. If the reaction is more severe and prolonged, some of the activated macrophages will have fused together to form large cells containing several nuclei. An accumulation of activated macrophages of this sort is termed a granuloma. Immunity to a number of other diseases (for example, leprosy, leishmaniasis, coccidiosis, and brucellosis) also can be gauged by the presence or absence of a delayed reaction to a test injection of the appropriate antigen. In all these cases, the test antigen provokes only a transitory response when the test is positive and, of course, no response at all when the test is negative.

Formation of granulomas

The same cell-mediated mechanisms are elicited by an actual infection with the living microbes, in which case the inflammatory response continues and the ensuing tissue damage and granuloma formation can cause serious damage. Moreover, in an actual infection, the microbes are often present inside the macrophages and are not necessarily localized in the skin. Large granulomas develop when the stimulus persists, especially if undegradable particulate materials are present and several macrophages, all attempting to ingest the same material, have fused their cell membranes to one another. The macrophages continue to secrete enzymes capable of breaking down proteins, and the normal structure of tissues in their neighbourhood becomes distorted. Although granuloma formation may be an effective method the immune system employs to sequester indigestible materials (whether or not of microbial origin) from the rest of the body, the harm inflicted by this immune mechanism may be much more serious than the damage caused by the infectious organisms. This is the case in such diseases as pulmonary tuberculosis and schistosomiasis and in certain fungal infections that become established within the body tissues rather than at their surface.

Contact hypersensitivity and dermatitis. In contact hypersensitivity, inflammation occurs when the sensitizing chemical comes in contact with the skin surface. The chemical interacts with proteins of the body, altering them so that they appear foreign to the immune system. A variety of chemicals can cause this type of reaction. They include various drugs, excretions from certain plants, metals such as chromium, nickel, and mercury, and industrial products such as hair dyes, varnish, cosmetics, and resins. All these diverse substances are similar in that they can diffuse through the skin. One of the best-known examples of a plant that can provoke a contact hypersensitivity reaction

is poison ivy (*Toxicodendron radicans*), found throughout North America. It secretes an oil called urushiol, which is also produced by poison oak (*T. diversilobum*), the poison primrose (*Primula obconica*), and the lacquer tree (*T. vernicifluum*). When urushiol comes in contact with the skin, it initiates the contact hypersensitivity reaction.

As sensitizing chemicals diffuse into the skin, they react with some proteins of the body, changing the antigenic properties of the protein. The chemical can interact with proteins located in both the outer horny layer of the skin (dermis) and the underlying tissue (epidermis). Some of the epidermal protein complexes migrate to the draining lymph nodes, where they stimulate T cells responsive to the newly formed antigen to multiply. When the T cells leave the nodes to enter the bloodstream, they can travel back to the site where the chemical entered the body. If some of the sensitizing substance remains there, it can reactivate the T cells, inducing a recurrence of inflammation.

Contact dermatitis

The clinical result is contact dermatitis, which can persist for many days or weeks. Treatment is by local application of corticosteroids, which greatly diminish lymphocyte infiltration, and by avoidance of further contact with the sensitizing agent.

Although delayed hypersensitivity can be a nuisance when it produces skin allergies, it is an important part of the immune defense against intracellular parasites, and it may also play a role in the containment of some tumours.

Autoimmune disorders

The mechanism by which the enormous diversity of B and T cells is generated is a random process that inevitably gives rise to some receptors that recognize the body's own constituents as foreign. Lymphocytes bearing such self-reactive receptors, however, are eliminated or rendered impotent by several different mechanisms, so that the immune system does not normally generate significant amounts of antibodies or T cells that are reactive with the body's components (self antigens). Nevertheless, an immune response to self, called autoimmunity, can occur, and some of the ways that self-directed immune responses cause damage have been mentioned above in *Allergies*.

Understanding and identifying autoimmune disorders is difficult given that all humans have many self-reactive antibodies in the blood but most show no sign of disease. Consequently the identification of autoantibodies is not a sufficient diagnostic tool for determining the presence of an autoimmune disorder. There is a difference between an autoimmune response and disease: in the former case the autoantibodies do not cause dysfunction, but in the latter case they do.

BASIC PROCESSES UNDERLYING AUTOIMMUNITY

Immunologists cannot always explain why the mechanisms that normally prevent the development of autoimmunity have failed in a particular autoimmune disorder. They have, however, advanced a number of explanations for such failures.

Alteration of self antigens. Various mechanisms can alter self components so that they seem foreign to the immune system. New antigenic determinants can be attached to self proteins, or the shape of a self antigen can shift—for a variety of reasons—so that previously unresponsive helper T cells are stimulated and can cooperate with pre-existing B cells to secrete autoantibodies. Alteration of the shape of a self protein has been shown to occur in experimental animals and is the most probable explanation for the production of the rheumatoid factors that are characteristic of rheumatoid arthritis. Infectious organisms also can alter self antigens, which may explain why viral infection of specialized cells—such as those in the pancreas that secrete insulin or those in the thyroid gland that make thyroid hormones—often precedes the development of autoantibodies against the cells themselves and against their hormonal products.

Release of sequestered self antigens. Intracellular antigens and antigens found on tissues that are not in contact with the circulation normally are segregated effectively from the immune system. Thus, they may be regarded as foreign if they are released into the circulation as a result of tissue destruction caused by trauma or infection. After sudden damage to the heart, for example, antibodies against heart muscle membranes regularly appear in the blood.

Cross-reaction with foreign antigens. This mechanism comes into play when an infectious agent produces antigens so similar to those on normal tissue cells that the antibodies stimulated to react against the foreign antigen also recognize the similar self antigen; hence, the two antigens are said to be cross-reactive. Autoantibodies stimulated by external antigens in this way tend to persist. For example, the streptococci that cause rheumatic fever make antigens that are cross-reactive with those on heart muscle membranes, and the antibodies that react with the bacteria also bind to the heart muscle membrane and cause damage to the heart. Another instance of an autoimmune disorder that arises from cross-reactivity is Chagas' disease. The trypanosomes that cause the disease make antigens that are cross-reactive with antigens on the surface of the specialized nerve cells that regulate the orderly contraction of muscles in the bowel. Antibodies directed against the trypanosomes also interact with these nerve cells and disrupt normal bowel functioning.

Genetic factors. Several autoimmune diseases clearly run in families. Careful studies (for example, those comparing the incidence in identical twins with that in fraternal twins) have shown that the increased incidence of such autoimmune diseases cannot be explained by environmental factors. Rather, it stems from a genetic defect that is passed from one generation to the next. Such disorders include Graves' disease, Hashimoto's disease, autoimmune gastritis (including pernicious anemia), type I (insulin-dependent) diabetes mellitus, and Addison's disease. These diseases are more common in persons who bear particular MHC antigens on their cells. The possession of these antigens does not imply that a person will contract such diseases, only that he or she is more likely to do so. Researchers generally agree that the interaction of many genes is needed before a person develops such autoimmune diseases. For example, type I diabetes is believed to result from at least 14 genes.

Inherited autoimmune diseases

Another interesting feature that appears to relate to the inheritance of autoimmune disorders is gender. Most human autoimmune diseases afflict far more women than men. Women are affected more often than men with most of the better-known disorders, including myasthenia gravis, systemic lupus erythematosis, Graves' disease, rheumatoid arthritis, and Hashimoto's disease. The reason for this is not fully understood, but researchers think it probably is related to hormonal effects on immune responses.

EXAMPLES OF AUTOIMMUNE DISORDERS

The spectrum of autoimmune disorders is wide, ranging from those that involve a single organ to others that affect several different organs as a secondary consequence of the presence of immune complexes in the circulation. It is not possible in this article to discuss them all. The following disorders have been chosen to illustrate some of the very different complications that can arise from autoimmunity.

Autoimmune diseases of the thyroid gland. Hashimoto's disease and Graves' disease are two of the most common autoimmune disorders of the thyroid gland, the hormone-secreting organ (located in the throat near the larynx) that plays an important role in the development and maturation of all vertebrates. The thyroid is composed of closed sacs (follicles) lined with specialized thyroid cells. These cells secrete thyroglobulin, a large protein that acts as a storage molecule from which thyroid hormones are made and released into the blood. The rate at which this occurs is regulated by thyroid-stimulating hormone (TSH), which activates the thyroid cells by combining with TSH receptors found on the thyroid cell membrane. Hashimoto's disease involves swelling of the gland (a condition called goiter) and a loss of thyroid hormone production (hypothyroidism). The autoimmune process underlying this disorder is thought to be instigated by helper T cells that react with thyroid antigens, although the mechanism is not completely understood. Once activated, the self-reactive T

Hashimoto's disease and Graves' disease

cells stimulate B cells to secrete antibodies against several target antigens, including thyroglobulin.

Graves' disease is a type of overactive thyroid disease (hyperthyroidism) involving excess production and secretion of thyroid hormones. The disease arises with the development of antibodies that are directed against the TSH receptor on the thyroid cells and that can mimic the action of TSH. When bound to the receptor, the antibodies stimulate excessive secretion of thyroid hormones.

In both Hashimoto's disease and Graves' disease, the thyroid gland becomes infiltrated with lymphocytes and is partially destroyed. If the gland is completely destroyed, a condition called myxedema may ensue, involving a swelling of tissues, especially those around the face.

Autoimmune hemolytic anemia. A number of autoimmune disorders are grouped under the rubric autoimmune hemolytic anemia. All result from the formation of autoantibodies against red blood cells, an event that can lead to hemolysis (destruction of red blood cells). The autoantibodies sometimes appear after infection with the bacterium *Mycoplasma pneumoniae,* a rather uncommon cause of pneumonia. In that case the autoantibodies are directed against certain antigens that are present on red cells, and they are probably induced by a similar antigen in the microbes (an example of the cross-reaction of antigens). Autoantibodies directed against a different antigen of red blood cells are often produced in persons who have been taking the antihypertensive medication alpha methyldopa for several months; the reason for autoantibody development in such cases is unknown. Other drugs, such as quinine, sulfonamides, or even penicillin, very occasionally cause hemolytic anemia. In such cases it is thought that the drug acts as a hapten—that is, it becomes bound to a protein on the surface of red blood cells, and the complex becomes immunogenic.

The autoantibodies that form against red blood cells are categorized into two groups on the basis of their physical properties. Autoantibodies that bind optimally to red blood cells at 37° C (98.6° F) are categorized as warm-reacting. Warm-reacting autoantibodies belong primarily to the IgG class and cause about 80 percent of all cases of autoimmune hemolytic anemia. Autoantibodies that attach to red blood cells only when the temperature is below 37° C are called cold-reacting. They belong primarily to the IgM class. Cold-reacting autoantibodies are efficient at activating the complement system and causing the cell to which they are bound to be destroyed. Nevertheless, as long as the body temperature remains at 37° C, cold-reacting autoantibodies dissociate from the cell, and hemolysis is not severe. However, when limbs and skin are exposed to the cold for long periods of time, the temperature of circulating blood can be lowered, allowing cold-reacting autoantibodies to go to work. Infection with *M. pneumoniae* is met by cold-reacting antibodies.

Pernicious anemia and autoimmune gastritis. Pernicious anemia stems from a failure to absorb vitamin B_{12} (cobalamin), which is necessary for the proper maturation of red blood cells. It is characteristically accompanied by a failure to secrete hydrochloric acid in the stomach (achlorhydria) and is in fact a symptom of severe autoimmune gastritis. To be absorbed by the small intestine, dietary vitamin B_{12} must form a complex with intrinsic factor, a protein secreted by the parietal cells in the stomach lining. Pernicious anemia results when autoantibodies against intrinsic factor bind to it, preventing it from binding to vitamin B_{12} and thus preventing the vitamin from being absorbed into the body. The autoantibodies also destroy the acid-secreting parietal cells, which leads to autoimmune gastritis.

Rheumatoid arthritis. Rheumatoid arthritis is a chronic inflammatory disease that affects connective tissues throughout the body, particularly the synovial membranes that line the peripheral joints. Rheumatoid arthritis is one of the most common autoimmune diseases. Its cause is not known, but a variety of altered immune mechanisms probably contribute to the disorder, especially in more severe cases.

One theory suggests that the inflammatory process of the disease is initiated by autoimmune reactions that involve one or more autoantibodies, referred to collectively as rheumatoid factor. The autoantibodies react with the tail region of the Y-shaped IgG molecule—in other words, rheumatoid factor is anti-IgG antibodies. Immune complexes form between rheumatoid factor and IgG and apparently are deposited in the synovial membrane of joints. The deposition triggers a type III hypersensitivity reaction, activating complement and attracting granulocytes, which causes inflammation and pain in the joints. The granulocytes release enzymes that break down cartilage and collagen in the joints, and this eventually can destroy the smooth joint surface that is needed for ease of movement. If immune complexes in the blood are not effectively removed by the liver and spleen, they can produce systemic effects similar to those precipitated by serum sickness.

Rheumatoid factor

The devastating effects of rheumatoid arthritis also have been seen in patients, especially younger ones, in whom no rheumatoid factor is detected, and thus other mechanisms of initiation of the disorder probably exist.

Systemic lupus erythematosus. Systemic lupus erythematosus (SLE) is a syndrome characterized by organ damage that results from the deposition of immune complexes. The immune complexes form when autoantibodies are made against the nucleic acids and protein constituents of the nucleus of cells. Such autoantibodies, called antinuclear antibodies, do not attack healthy cells, since the nucleus lies within the cell and is not accessible to antibodies. Antigen-antibody complexes form only after the nuclear contents of a cell are released into the bloodstream during the normal course of cell death or as a result of inflammation. The resultant immune complexes are deposited in tissues, causing injury. Certain organs are more commonly involved than others, including the kidneys, joints, skin, heart, and serous membranes around the lungs.

Multiple sclerosis. Multiple sclerosis is an autoimmune disease that results in the gradual destruction of the myelin sheath that surrounds nerve fibres. It is characterized by progressive degeneration of nerve function, interjected with periods of apparent remission. The cerebrospinal fluid of persons with multiple sclerosis contains large numbers of antibodies directed against myelin basic protein and perhaps other brain proteins. Infiltrating lymphocytes and macrophages may exacerbate the destructive response. The reason the immune system launches an attack against myelin is unknown, but several viruses have been suggested as initiators of the response. A genetic tendency toward the disease has been noted; susceptibility to the disorder is indicated by the presence of the major histocompatibility complex (MHC) genes, which produce proteins found on the surface of B cells and some T cells.

Type I (insulin-dependent) diabetes mellitus. Type I diabetes mellitus is the autoimmune form of diabetes and often arises in childhood. It is caused by the destruction of cells of the pancreatic tissue called the islets of Langerhans. Those cells normally produce insulin, the hormone that helps regulate glucose levels in the blood. Individuals with type I diabetes have high blood glucose levels that result from a lack of insulin. Dysfunction of islet cells is caused by the production of cytotoxic T cells or autoantibodies that have formed against them. Although the initiating cause of this autoimmune response is unknown, there is a genetic tendency toward the disease, which also involves class II MHC genes. It can be treated with injections of insulin; however, even when treated, type I diabetes may eventually lead to kidney failure, blindness, or serious circulation difficulties within the extremities.

Other autoimmune disorders. Mechanisms similar to those that produce autoimmune hemolytic anemia can result in the formation of antibodies against granulocytes and platelets, although autoimmune attacks against these blood cells occur less frequently. Antibodies against other types of cells occur in a number of autoimmune diseases, and those self-reactive responses may be primarily responsible for the damage incurred. In myasthenia gravis, a disease characterized by muscle weakness, autoantibodies react against receptors on muscle cells. Normally the receptors bind to acetylcholine, a neurotransmitter released from nerve endings. When acetylcholine binds to an acetylcholine receptor on the surface of muscle cells, it stimulates the muscle to contract. The autoantibodies in

Myasthenia gravis

myasthenia gravis bind to the acetylcholine receptors without activating them. The antibodies prevent muscle contraction either by blocking acetylcholine from binding to its receptor or by destroying the receptors outright. This renders the muscle less responsive to acetylcholine and ultimately weakens muscle contraction.

Cancers of the lymphocytes

Tumours arising from lymphocytes are given various names: they are called leukemias if the cancer cells are present in large numbers in the blood, lymphomas if they are mainly concentrated in lymphoid tissues, and myelomas if they are B-cell tumours that secrete large amounts of immunoglobulin. The following sections describe how cancers of the lymphocytes arise and how immunological techniques are being used to determine the prognosis and treatment of B- and T-cell tumours.

GENETIC CAUSES

Most cancers result from a series of random genetic accidents, or mutations, that occur to genes involved in controlling cell growth. One general group of genes implicated in cancer initiation and growth are called oncogenes. The unaltered, healthy form of an oncogene is called a proto-oncogene. Proto-oncogenes stimulate cell growth in a controlled manner that involves the interplay of a number of other genes. However, should a proto-oncogene become mutated in some way, it may become hyperactive, leading to uncontrolled cellular proliferation and the exaggeration of some normal cellular activities. A proto-oncogene can become mutated in a number of ways. According to one mechanism, called chromosomal translocation, part of one chromosome is severed from its normal position and reattached (translocated) onto another chromosome. If a proto-oncogene appears on the piece of the chromosome that is moved, it may be separated from the region that normally regulates it. In this manner the proto-oncogene becomes unregulated and turns into an oncogene. Chromosomal translocation of proto-oncogenes is involved in a number of B-cell tumours, including Burkitt's lymphoma and chronic myelogenous leukemia. T-cell leukemia also results from a chromosomal translocation.

MALIGNANT TRANSFORMATION OF LYMPHOCYTES

At any stage in its development, from stem cell to mature form, a lymphocyte may undergo malignant (cancerous) transformation. The transformed cell is no longer constrained by the processes that regulate normal development, and it proliferates to produce a large number of identical cells that make up the tumour. These cells retain the characteristics of the transformed cell's particular developmental stage, and because of this cancers can be distinguished according to the stage at which transformation took place. For example, B cells that become cancerous in the early stages of development give rise to such conditions as chronic myelogenous leukemia and acute lymphocytic leukemia, whereas malignant transformation of late-stage B cells—*i.e.*, plasma cells—can result in multiple myeloma. Regardless of what stage of the cell becomes cancerous, malignant cells outgrow and displace other cells that continue to develop normally.

Leukemia

TREATMENT

Both T and B cells have surface antigens that are characteristic of different stages in their life cycle, and antibodies have been prepared that identify the antigens. Knowledge of the specific type and stage of maturation of the tumour cells helps physicians determine the prognosis and course of treatment for the patient. This information is important because different types of tumours respond to different therapies and also because the chances of effecting a cure vary from type to type. Advances in drug treatments have dramatically improved the outlook for children with acute lymphoblastic leukemia, the most prevalent of the childhood leukemias. Similarly, most cases of Hodgkin's disease, a common type of lymphoma that mainly strikes adults, can be cured by drugs, radiation, or a combination of both. Myelomas primarily arise in older individuals. These tumours grow fairly slowly and are usually diagnosed by virtue of the characteristic immunoglobulin they secrete. However, this immunoglobulin may be produced in such large amounts that it causes secondary damage such as kidney failure.

BIBLIOGRAPHY

Textbooks: NORMAN A. STAINES, JONATHAN BROSTOFF, and KEITH JAMES, *Introducing Immunology,* 2nd ed. (1993), with extensive illustrations; J.H.L. PLAYFAIR, *Immunology at a Glance,* 6th ed. (1996); ELI BENJAMINI, GEOFFREY SUNSHINE, and SIDNEY LESKOWITZ, *Immunology: A Short Course,* 3rd ed. (1996), well-constructed and very readable; KLAUS D. ELGERT, *Immunology: Understanding the Immune System* (1996), covering the basic principles; and LESLEY-JANE EALES, *Immunology for Life Scientists* (1997), also on the basic principles. DAVID MALE, *Immunology: An Illustrated Outline,* 2nd ed. (1991); and IVAN ROITT, JONATHAN BROSTOFF, and DAVID MALE (eds.), *Immunology,* 5th ed. (1998), teaches immunology through extensive figures and diagrams.

Additional texts: WILLIAM E. PAUL (ed.), *Fundamental Immunology,* 3rd ed. (1993), requires a solid foundation in biological principles and some knowledge of immunology. Also recommended are EMIL R. UNANUE and BARUJ BENACERRAF, *Textbook of Immunology,* 2nd ed. (1984); and JOSEPH A. BELLANTI (ed.), *Immunology: Basic Processes,* 2nd ed. (1985). JULIUS M. CRUSE and ROBERT E. LEWIS, *Illustrated Dictionary of Immunology* (1995), a comprehensive work, is useful for experienced immunologists and general readers.

History: H.J. PARISH, *Victory with Vaccines: The Story of Immunization* (1968); and J.H. HUMPHREY and R.G. WHITE, *Immunology for Students of Medicine,* 3rd ed. (1970), especially the introductory chapter, are worth consulting. JAN KLEIN, *Immunology: The Science of Self-Nonself Discrimination* (1982), explains clearly how and why principles were discovered. ARTHUR M. SILVERSTEIN, *A History of Immunology* (1989), an advanced text, analyzes the history of immunology from 1720 to 1970, thoroughly discussing the discovery of immunological principles and the scientists involved. LESLIE BRENT, *A History of Transplantation Immunology* (1997), also an advanced text, examines the scientific discoveries that developed this field of immunology.

Immune system disorders: LLOYD J. OLD, "Immunotherapy for Cancer," in *Scientific American,* 275(9):136–143 (September 1996), explores the use of the human immune system in combating cancer. Clinical aspects are covered by G.L. ASHERSON and A.D.B. WEBSTER, *Diagnosis and Treatment of Immunodeficiency Diseases* (1980); MAX SAMTER (ed.), *Immunological Diseases,* 4th ed., 2 vol. (1988); and P.J. LACHMANN *et al.,* (eds.), *Clinical Aspects of Immunology,* 5th ed., 3 vol. (1993).

Specialized topics: H. HUGH FUDENBERG *et al., Basic Immunogenetics,* 3rd ed. (1984), covers genetic aspects. Immunochemistry is treated in ELVIN A. KABAT, *Structural Concepts in Immunology and Immunochemistry,* 2nd ed. (1976); and I.E. GLYNN and M.W. STEWARD, *Immunochemistry: An Advanced Textbook* (1977). EDWIN L. COOPER, *General Immunology* (1982), describes the evolution of the immune system.

(J.H.Hy./S.S.P.)